AARHUS
UNIVERSITY
HEALTH

A statistical perspective on association studies of psychiatric disorders: genetic effects of single-markers, haplotypes, gene-environment interactions and gene-gene interactions

PhD dissertation

Leslie Foldager

Health
Aarhus University
2014

A statistical perspective on association studies of psychiatric disorders: genetic effects of single-markers, haplotypes, gene-environment interactions and gene-gene interactions

PhD dissertation

Leslie Foldager

Health
Aarhus University
Department of Clinical Medicine
Translational Neuropsychiatry Unit

PhD student, Statistician, MSc

Translational Neuropsychiatry Unit (TNU)
Department of Clinical Medicine, Health
Aarhus University, Risskov, Denmark (1 Apr 2013 –)

Centre for Psychiatric Research
Aarhus University Hospital, Risskov
Central Denmark Region, Denmark (10 Nov 2002 – 31 Mar 2013)

Bioinformatics Research Centre (BiRC)
Department of Computer Science, Science and Technology
Aarhus University, Aarhus, Denmark (1 Mar 2008 –)

*i*PSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research
Aarhus and Copenhagen, Denmark (1 Mar 2012 –)

*i*SEQ, Centre for Integrative Sequencing, Aarhus University, Denmark (1 Dec 2012 –)

Supervisors

Anders D. Børglum (main supervisor), Professor, PhD, MD
Department of Biomedicine, Aarhus University, Aarhus, Denmark

Ole Mors (project supervisor), Professor, PhD, MD
Department of Clinical Medicine
Aarhus University and Aarhus University Hospital, Risskov, Denmark

Carsten Wiuf, Professor, PhD, MSc
Department of Mathematical Science, University of Copenhagen, Copenhagen, Denmark

Evaluation committee and defence

Palle Villesen Fredsted (chairman of committee), Associate Professor, PhD, MSc
Bioinformatics Research Centre and Department of Clinical Medicine
Aarhus University, Aarhus, Denmark

Andrew McQuillin, Senior Lecturer, PhD, MSc
Molecular Psychiatry Laboratory, Mental Health Sciences Unit
Faculty of Brain Sciences, University College London, United Kingdom

Claus Thorn Ekstrøm, Professor, PhD, MSc
Section of Biostatistics, Department of Public Health
University of Copenhagen, Copenhagen, Denmark

Gregers Wegener (chairman of defence), Professor, DMSc, PhD, MD
Translational Neuropsychiatry Unit, Department of Clinical Medicine
Health, Aarhus University, Risskov, Denmark

Date of defence: June 11, 2014

Contents

Preface

Worldwide collaborations in large consortia are showing promising results concerning identification of genetic loci contributing to the risk of complex diseases. This collaboration was initially mainly based on the bulk of genome-wide association studies (GWAS) that have been performed for many diseases over the last 5–10 years. Now that the hurdle of collaborating rather than competing has been crossed, it appears that these consortia facilitate new studies of much larger size than we would even have dreamed of just few years ago. The studies carried out under The Lundbeck Foundation Initiative for Integrative Psychiatric Research, *i*PSYCH, is an excellent example of this. At the same time state-of-the-art for statistical genetics changes at a staggering speed, and psychiatric genetics is one of the cradles for methodological developments, maybe due to the very complex nature of the psychiatric disorders investigated. So it is definitely a very exciting time to be involved in psychiatric genetics. Technological developments imply ever larger amounts of genetic and genomic data, and consequently the capacity and statistical methodology to analyse and interpret the data constantly lag behind. Nevertheless, all the more reason to keep on developing methods and investigate which are the better alternatives to current or older standards. The present dissertation is a contribution to this and was carried out as a part time study at Health, Aarhus University, from September 2007 to March 2014.

Outline

The dissertation was constructed from a selection of six manuscripts of which four have been published (paper 1, 2, 4 and 5). These four papers are as such presenting results from ordinary health science studies within the field of psychiatric genetics (case-control studies) but imply also various statistical issues that we have to consider and handle. A common thread is the awareness of checking for possible interactions both between genetic markers and between these markers and non-genetic factors. The phenotypes considered in these studies are schizophrenia (paper 1 and 5), bipolar disorder (paper 2 and 5), panic disorder (paper 2), and suicidal behaviour (paper 4). Though it may seem a rather incoherent way of doing research, most psychiatric disorders have overlapping symptoms and to some extent shared hypotheses about the aetiology. Certainly, the studies included in this thesis are not using genome-wide strategies, and the sample sizes are modest in view of today's standard. Nevertheless, we believe that focused genotyping and investigation of specific hypotheses or pathways will continue to be relevant even in the present GWAS and next generation sequencing (NGS) era of psychiatric genetics. The last two papers are different as they utilise computer simulations instead of real data. Paper 3 presents the initial steps (mainly data construction) of an ongoing simulation study aiming at comparing methods for gene-environment (G×E) interaction analysis. And finally, paper 6 introduces a new method that can be used to summarise a series of sequentially ordered stochastic variables, e.g. to aggregate p-values without a priori grouping.

The introduction contains background and aims both for the thesis and for each of the papers/studies. Material and methods contains a condensed and complementary description of samples, selection of genes and genetic markers, and statistical methods. Further details including laboratory methods for DNA extraction, genotyping and serum determination can be found in the papers. A separate chapter is devoted to statistical methods, extending the descriptions given in the papers but also giving some details with a view towards perspectives and future plans. After the methods chapters, the results from the studies are presented, and then follows a discussion chapter including conclusions and perspectives. Lastly, the included manuscripts are reproduced, followed by appendices with some further technical details.

To be able to differentiate between cross references to the manuscripts and cross references to the thesis, references to the manuscripts will start with a capital letter (e.g. *Figure*) whereas references to the thesis will start with lowercase letters (e.g. *figure*). All referenced links to web pages were accessed and working on March 31, 2014.

Manuscripts

The thesis is based on the following manuscripts:

1. **Leslie Foldager**, Rudi Steffensen, Steffen Thiel, Thomas Damm Als, Hans Jørgen Nielsen, Merete Nordentoft, Preben Bo Mortensen, Ole Mors and Jens Christian Jensenius. *MBL and MASP-2 concentrations in serum and MBL2 promoter polymorphisms are associated to schizophrenia*. Acta Neuropsychiatrica 2012; **24**(4): 199–207.

2. **Leslie Foldager**, Ole Köhler, Rudi Steffensen, Steffen Thiel, Ann Suhl Kristensen, Jens Christian Jensenius and Ole Mors. *Bipolar and panic disorders may be associated with hereditary defects in the innate immune system*. Journal of Affective Disorders 2014; **164**: 148–154.

3. **Leslie Foldager**, Thomas Damm Als and Jakob Grove. *Comparison of methods for genome-wide gene-environment interaction analysis*. Manuscript in preparation.

4. Henriette Nørmølle Buttenschøn*, Tracey J. Flint*, **Leslie Foldager**, Ping Qin, Søren Christoffersen, Nikolaj F. Hansen, Ingrid Bayer Kristensen, Preben Bo Mortensen, Anders D. Børglum and Ole Mors. *An association study of suicide and candidate genes in the serotonergic system*. Journal of Affective Disorders, 2013; **148**(2–3): 291–298.

5. Henriette Nørmølle Buttenschøn*, **Leslie Foldager***, Tracey J. Flint, Inger Marie L. Olsen, Thomas Deleuran, Mette Nyegaard, Mette Mejlby Hansen, Pekka Kallunki, Kenneth Vielsted Christensen, Douglas H. Blackwood, Walter J. Muir, Steen E. Straarup, Thomas Damm Als, Merete Nordentoft, Anders D. Børglum and Ole Mors. *Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3*. Psychiatric Genetics 2010; **20**(3): 93–101.

6. Carsten Wiuf*, Jonatan Schaumburg-Müller Pallesen*, **Leslie Foldager** and Jakob Grove. *Landscape: A simple method to aggregate p-values and other stochastic variables without a priory grouping*. Manuscript in preparation.

*) Authors contributing equally to the study.

Acknowledgements

First of all I am very thankful to my main supervisor Professor Anders Børglum and my project supervisor Professor Ole Mors who I have been working with since Nov 2002, i.e. my entire carrier in psychiatric research at the Psychiatric Hospital in Risskov. Actually my workplace has had many names during these years: Department of Demography (Institute of Psychiatric Demography), Centre for Basic Psychiatric Research, Centre for Psychiatric Research and now Translational Neuropsychiatry Unit. I'm in debt to the very many colleagues that have been coming and going over the years and of course to my current and former department/centre/unit heads for their support and collaboration.

Specially warm thanks to my colleague and current personnel manager Dorthe Eggertsen for her great support over the years. Impressively enthusiastic and always fights for the best conditions for her employees. We have had many discussions over the years not so much about research but more about rules, regulations, hiring, firing, organisational changes and so forth but always in an orderly manner even though she represented the workplace and I the employees and trade union.

I am also grateful to my supervisor Professor Carsten Wiuf who first of all paved the way for my long affiliation with Bioinformatics Research Centre (BiRC) starting 6 years ago in the old red officers building at Høegh Guldbergsgade, which is now history. In fact my affiliation with BiRC has been lasting so long that the permanent staff at BiRC have categorised me as a *hang-around* ... maybe even a *prospect*. I hope to keep hanging around at BiRC as it is a great and inspiring workplace. Maybe I could rise in rank and get a badge?

During my PhD study I have had the pleasure of working with so many nice people at the Translational Neuropsychiatry Unit (TNU), Bioinformatics Research Centre (BiRC), National Centre for Register-based Research (NCRR), Research Department P in Risskov, and Anders Børglums lab at Department of Biomedicine. I'm thankful to you all for inspiring talks, collaboration, journal clubs and seminars, good laughs during breaks, social events and so forth. I'm also thankful to my many other collaborators and fellows from around Denmark and abroad. Some are mentioned as co-authors of the manuscripts but many more would deserve being mentioned. The full list of names would be extremely long but nobody named, nobody forgotten. I do want to name a few though: Henriette Nørmølle Buttenschøn with whom I have had the pleasure of working and collaborating with ever since 2002 in "the genetics group" (Ole Mors lab) in Risskov; Thomas Damm Als a good colleague and partner in crime (social events) throughout all my years in Risskov even though he went on a long "fishing expedition" to Silkeborg; Jakob Grove my office fellow at BiRC and fourth supervisor (not mentioned but probably should have been); my office fellow at TNU Noomi Gregersen, and Nicklas Heine Staunstrup and Marit Nyholm Nielsen the last three current members of "the genetics group"; Mette Nyegaard for good discussions, collaboration, and joyful moments at social events and karaoke bars around the world; and many more could be mentioned. My appreciation to Karen Jul Madsen and Hella Storgaard Kastbjerg for proofreading parts of the dissertation on short notice before printing.

I also want to thank my family, friends, sport mates, choir, vocal groups and music bands for all the joy in my spare time. Last but not least a deep felt thanks to my wife Gitte and my children Mathias, Anders and Nicoline for enduring the many hours that I have spent in front of the computer—not least the last three months of 24–7 workload.

Leslie Foldager
June 1, 2014

Financial support

The Health faculty of Aarhus University covered study fees. Salary and most of the other running expenses were covered by the workplace, Center for Psychiatric Research now Translational Neuropsychiatry Unit. Translational Neuropsychiatry Unit covered the printing of the dissertation. A grant from Central Region Denmark covered half of the salary for a 5 month period of leave from other job duties. A grant from *Trepartsmidler til kompetenceudvikling* (Central Denmark Region) supported the participation in Summer Institute in Statistical Genetics (SISG) 2010. Grants from *Fonden til psykiatriens fremme* supported participation in SISG 2010 and World Congress of Psychiatric Genetics (WCPG) 2007. Grants from *Ole J. Rafaelsens Fond* supported participation in WCPG 2007 and 2010. The Lundbeck Foundation Initiative for Integrative Psychiatric Research, *i*PSYCH, is funded by the Lundbeck Foundation, Denmark. The Lundbeck Foundation had no involvement in any aspect of the study.

English summary

Gene-gene (G×G) and gene-environment (G×E) interactions likely play an important role in the aetiology of complex diseases like psychiatric disorders. Thus, we aim at investigating methodological aspects of and apply methods from statistical genetics taking interactions into account. In addition we consider issues concerning detection limits of continuous traits, single-marker tests, analysis of sex chromosomes, and accumulation of signals. Disorders investigated include schizophrenia, bipolar disorder, panic disorder, and suicidal behaviour. In addition to this, we use computer simulations.

Papers 1 and 2 were motivated by the hypothesis that defects of the immune system may increase risk of psychiatric disorders. We consider two components from the lectin pathway of activation: mannan-binding lectin (MBL) and MBL-associated serine protease-2 (MASP-2) via continuous traits (protein level), dichotomous trait (disease status) as well as genetic markers including G×G interactions. We use Tobit regression to handle data below the detection limit of MBL.

The involvement of the immune system may also be less direct as seen by the findings how infections impact disorders, e.g. via interaction between genes and maternal infection by virus. Paper 3 presents the initial steps (mainly data construction) of an ongoing simulation study aiming at guiding decisions by comparing methods for G×E interaction analysis including both traditional two-step logistic regression, exhaustive searches using efficient algorithms, and data mining or machine learning methods like model-based multifactor dimensionality reduction (MB-MDR) and logic regression with feature selection (logicFS).

The analysis of sex chromosomes may require different approaches than those commonly used for autosomes. In paper 4 we include a marker from the X chromosome and discuss how to analyse with and without the assumption of inactivation of one of the female X chromosomes early in development. In addition this paper includes analysis of the interaction between genetic markers and age and sex.

Haplotype analysis and other multilocus approaches may increase the power to detect disease association but introduce also the problem of determining the gametic phase. In papers 1 and 2 we analyse multilocus genotypes and haplotypes but assuming known phase as linkage disequilibrium (LD) implies only few haplotypes to be commonly observed using these markers. However, the validity of the identified haplotypes is also checked by inferring phased haplotypes from genotypes. Haplotype analysis is also used in paper 5 which is otherwise an example of a focused approach to narrow down a previously found signal to search for more precise positions of disease causing mutations and functional implications.

In stark contrast to such a focused approach stand genome-wide studies (GWAS). Here it is truly important to address the enormous increase in type I error introduced when performing hundreds of thousands or even millions of statistical tests. The commonly accepted genome-wide threshold for single-marker association tests has become 5e-8 but searching for interactions genome-wide results in drastically many more tests and thus the need of an even lower p-value threshold. Lowering the threshold comes at the unfortunate but inevitable expense of increasing the probability of type II errors and thus lowering the power to detect association. Statistical procedures where the test statistics initially are grouped according to some criteria, e.g. by candidate regions or functional pathways, may be one way to decrease the number of tests instead of lowering the threshold for significance. Yet, in paper 6 we propose the *Landscape* method to summarise a series of sequentially ordered test values without the need of more or less arbitrary prior grouping.

Danish summary

Gen-gen (G×G) og gen-miljø (G×E) interaktioner spiller sandsynligvis en ætiologisk rolle for komplekse sygdomme som f.eks. psykiske lidelser. Med det in mente er formålet derfor at undersøge metodiske aspekter samt anvende statistisk genetiske metoder, som kan håndtere disse interaktioner. Derudover ser vi også på problemstillinger som håndteringen af detektionsgrænser for kontinuerte målinger, test af enkelt-markører, analyse af kønskromosomer samt akkumulering af signaler. Afhandlingen inkluderer grupper af patienter med skizofreni, bipolar sygdom, panikangst og selvmordsadfærd. Derudover benyttes computer simuleringer.

Motiveringen for Artikel 1 og 2 var hypotesen om at defekter i immunsystemet kan øge risikoen for psykiske lidelser. Vi undersøger her to komponenter fra lektin-vejen til aktivering af komplementsystemet, mannanbindende lektin (MBL) og MBL-associeret serinprotease-2 (MASP-2). Disse komponenter undersøges via kontinuerte træk, dikotomt respons (sygdomsstatus) såvel som genetiske markører inklusiv G×G interaktioner. Til håndtering af MBL målinger under detektionsgrænsen benytter vi Tobit regression.

Involveringen af immunsystemet kan også ske mere indirekte som f.eks. via virusinfektioner hos moderen under graviditeten. I artikel 3 præsenteres indledende skridt (primært generering af data) til et igangværende simulationsstudie som har til formål at yde beslutningsstøtte til valg af metoder ved at sammenligne metoder til analyse af G×E interaktioner, herunder klassisk to-trins logistisk regression, effektive fuldt dækkende algoritmer samt ”data mining” og ”machine learning” metoder som f.eks. model-baseret multifaktor dimensionalitetsreduktion (MB-MDR) and logisk regression inklusiv såkaldt ”feature selection” (logicFS).

Analyse af kønskromosomer kræver eventuelt andre tilgange end de, der benyttes til autosomerne. I artikel 4 inkluderer vi en markør fra X kromosomet og vurderer på, hvordan man kan analysere med og uden en antagelse om inaktivering af det en kvindelige X kromosom tidligt i fostrets eller barnets udvikling. I denne artikel analyseres desuden for interaktioner mellem genetiske markører og køn og alder.

Haplotype analyse og andre multilokus metoder kan på den ene side øge styrken til at påvise sygdomsassociation, men medfører på den anden side behovet for at fastslå den gametiske fase. I artikel 1 og 2 benytter vi multilokus genotyper og haplotyper, men her sker det under antagelse af kendt fase, da der pga. koblingsuligevægt (LD) normalvist kun observeres et mindre antal haplotyper med disse markører. Vi tjekker dog også validiteten af de disse haplotyper ved at aflede fasede haplotyper ud fra genotyperne. Haplotype analyse benyttes også i artikel 5, som ellers primært er et eksempel på en analyse, der er målrettet mod at indsnævre et tidligere fund for derved at fastslå en mere præcis placering af sygdomsfremkaldende mutationer og funktionelle konsekvenser.

Helgenomsstudier (GWAS) står i skarp kontrast til sådanne fokuserede tilgange. Ved helgenomsstudier er det for alvor vigtig at håndtere den voldsomme øgning af type I fejl, der er en følge af at udføre hundredetusindvis eller måske endda millioner af statistiske test. Det er gængs at benytte 5e-8 som helgenoms tærskelværdi for enkelt-markør associationstestning. Denne grænse er dog langt fra lav nok, hvis der udføres en helgenomssøgning efter interaktioner. En lavere tærskel medfører imidlertid uvægerligt en øget risiko for at begå fejl af type II med deraf følgende lavere styrke. Statistiske procedurer som initialt grupperer testene efter nogle fastlagte kriterier, som f.eks. kandidatregioner eller funktionelle stier, er en mulig måde at reducere antallet af test uden at sænke tærsklen for signifikans. I artikel 6 indfører vi *Landskabsmetoden* som en måde at sammenfatte en fortløbende række af ordnede teststørrelser uden behovet for at lave en mere eller mindre tilfældig forudgående gruppering af testene.

Chapter 1

Introduction

In 2007 when the present PhD project was started, a major problem of many studies was a relatively week power due to small sample sizes in combination with the small genetic effects of individual susceptibility genes of complex disorders. This resulted both in few positive findings and a prominent lack of replication of the few findings reported. A recognized explanation seemed to be the contribution from multiple minor effect loci, gene-gene interactions between these (epistasis) together with non-genetic (environmental) effects and gene-environmental interactions. All of which demand larger sample sizes in well-designed studies that are evaluated with powerful statistical methods. In this respect the earliest genome-wide association studies (GWAS) from 2005 and the years to follow simply were too small.

We also participated with such an underpowered GWAS in the Danish Genomic Medicine for Schizophrenia (GEMS) project (Hollegaard et al., 2011) which included a little less than 900 patients with schizophrenia and equally many time-matched controls genotyped with the Illumina Infinium Human610-Quad bead chip. Punches from neonatal dried blood spot samples obtained from the Danish Newborn Screening Biobank (Norgaard-Pedersen et al., 2007) were used to obtain DNA for genotyping and material for testing for viral antibodies. The cases and controls were found by use of the Danish Psychiatric Central Register (Mors et al., 2011) and the Danish Civil Registration System (Pedersen et al., 2006). A few results have been published in relation to gene-environment interactions with viral antibodies from maternal infection, but most importantly the study revealed the enormous potential in utilising the Danish registries to link various sources of health and social information. In my view, the most import outcome from the GEMS study was the initiation in 2012 of the large Danish study of mental disorders, The Lundbeck Foundation Initiative for Integrated Psychiatric Research, *i*PSYCH[1].

Over the last couple of years it has been demonstrated that worldwide collaboration in large consortia increases the possibility to identify genetic loci contributing to the risk of complex diseases. This is facilitated via so-called mega-analyses of genome-wide association studies (GWAS), i.e. meta-analyses of tens of thousands or even hundreds of thousands of individuals. The exponential growth in the number of genome-wide studies seen since 2005 is now paying off (Visscher et al., 2012). Within the field of schizophrenia, this has convincingly been seen in mega-analyses from the Psychiatric Genomic Consortium (PGC). In the first wave (Ripke et al., 2011), seven loci passed the genome-wide significance threshold which is usually taken to be 5e-8. Five of these were new. In the second publication, Ripke et al. (2013) found twenty-two genome-wide significant loci and 13 of these were new. Between these publications, presentations at the

[1] `http://ipsych.au.dk`

World Congress of Psychiatric Genetics (WCPG) from PGC indicated the finding of more than 60 genome-wide significant sites for schizophrenia in 2012 and more than 100 in 2013. Large scale meta-analyses of bipolar disorder are also being conducted via PGC (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011) and confirm earlier suggestions as well as identify new risk genes.

Often the outcome (or trait) considered in genetic association studies is a dichotomous disease status (affected vs. unaffected, patients vs. healthy individuals, or simply stated: cases vs. controls) or maybe polytomous (e.g. healthy, mildly affected, severely affected, very severely affected) but in some studies quantitative traits measured at a continuous scale are also considered, e.g. volumetric measures of brain (Ahdidan et al., 2013) and serum concentration of brain-derived neurotrophic factor (BDNF) (Elfving et al., 2012), Vascular endothelial growth factor (VEGF) (Elfving et al., 2014), mannan-binding lectin (MBL) and MBL-associated serine protease-2 (MASP-2) (Foldager et al., 2012). In the present thesis, we consider four dichotomous phenotypes (schizophrenia, bipolar disorder, panic disorder and suicidal behaviour) and two quantitative traits: concentration of MBL and MASP-2 protein in serum. In principle these quantitative traits are measured on a continuous scale but a detection limit implies that the lowest MBL levels are unknown, i.e. the measure is interval censored. In addition a discretised version of MBL serum concentration is considered in terms of various categorisations, see subsection 2.2.1.

1.1 Background

1.1.1 Schizophrenia

Schizophrenia is among the most severe mental disorders and affects just over 7 per thousand adults (McGrath et al., 2008). From a social point of view, schizophrenia is among the most demanding illnesses with respect to expenses for treatment and transfer payment, and it is the 11th largest cause of years lived with disability (YLD), though ”only” the 106th most common condition (Murray et al., 2012, supplementary). According to the currently available tables (2000–2011) from WHO[2], schizophrenia is the 17th most common cause and accounts for 1.8% of the YLD worldwide (assessed 9 March, 2014).

The Global Burden of Disease (GBD) 2010 study (Murray et al., 2012) operates with two health states of schizophrenia: ”acute” and ”residual” (Ferrari et al., 2012). In the acute state, subjects are primarily burdened with so-called positive symptoms, perceiving things that most individuals normally do not experience (delusions, hallucinations and thought disorder) while individuals in the residual state are burdened mostly with negative symptoms (loss of interest, emotional deficits, asociality and lack of motivation). These states are not mutually exclusive and may co-occur. The proportion of studies surviving exclusion criteria for the GBD study was extremely low (6 of 188), and the results may therefore not be representative, but the indication was that approximately two-third of the cases were in the acute state.

Usually schizophrenia is thought to be a common disease (e.g. McClellan et al., 2007) with a lifetime risk of approximately 1% in the general population (McGuffin et al., 1995; Tamminga et al., 2005). But how prevalent does a disease (or disorder) have to be to be common? It seems difficult to actually track down a *common disease* definition but some clues of a *rare disease* definition exist—varying, however, much between continents. Rare diseases are also sometimes called orphan diseases and are object of legislative regulations and initiatives to

[2] `http://www.who.int/healthinfo/global_burden_disease/`

stimulate research and orphan drug development which may otherwise be of limited interest for the pharmaceutical industry for various reasons (Llinares, 2010; Tambuyzer, 2010; Forman et al., 2012). Due to these drug regulations a definition on prevalence limits for rare diseases can be devised. In USA a rare disease or disorder is defined as one affecting 200,000 Americans and thus corresponding to a point prevalence of approximately 1 per 1,500 (0.067%), in Japan the limit is 50,000 affected which corresponds to about 1 per 2,500 (0.04%), whereas EU operates with a prevalence threshold of 5 per 10,000 (i.e. 1 per 2,000 or 0.05%) (Llinares, 2010). Anyhow, schizophrenia is with estimated point prevalences around 45 per 10,000 or equivalently 0.45% (McGrath et al., 2008) much more common than any of these rare disease limits.

A systematic review of schizophrenia prevalence by Saha et al. (2005) indicates, however, that earlier prevalence estimates may be too high. In this review the median lifetime prevalence of schizophrenia was found to be 0.4% with no significant difference between genders, and they estimated the median lifetime morbid risk (LMR) to be 7.2 per 1,000. This is lower than the usual stated prevalence estimates between 0.5% and 1%, and e.g. the Diagnostic and Statistical Manual of Mental Disorders (DSM), Fourth Edition (DSM-IV) (American Psychiatric Association, 1994) reports the lifetime prevalence to be 1%. In a review McGrath et al. (2008) found that the incidence distribution is right skewed and varies between sites and between genders with a 1.4 male to female rate ratio of medians. Moreover, McGrath et al. (2008) estimated the median incidence to be 15.2 per 100,000, and the median lifetime prevalence estimate was 7.2 per 1,000, i.e. 0.7% and thus well in the 0.5–1% range. In a recent study covering the years 1995–2008 (Castagnini et al., 2013), we compared incidence and age of onset of acute and transient psychotic disorders (ATPDs) with schizophrenia and bipolar disorder for Danish males and females aged 15–64 years. Here the overall incidence rate of schizophrenia was 9.2 per 100,000 person-years but declining from 16.4 for subjects 15–24 years of age at onset to 2.8 for the 55–64 years age-band of onset. Castagnini et al. (2013) also observed a significantly higher incidence of schizophrenia for males with an overall male to female incidence rate ratio (IRR) of 2.0 with the gender difference peaking in the age-band 25–34 years (IRR=2.7) and disappearing totally for the oldest age-group (IRR=1.0). The median age of schizophrenia onset was 30.2 years, a bit lower for males (29.6 years) than females (31.7 years).

Common or not, patients and their relatives are being exposed to huge life strains and restrictions, and the mortality risk is markedly increased for persons with schizophrenia. A systematic review estimated the median all cause standardized mortality ratio (SMR) to be 2.5 (Saha et al., 2007) with the largest increases in the risk of dying observed for all unnatural causes and especially suicide: median SMRs of 7.5 and 12.9, respectively. McGrath et al. (2008) estimated the median all cause SMR to be 2.6, i.e. at the same level.

The causes of schizophrenia are largely unknown, it has a life-long course, prevention is not possible, and treatments are ineffective and burdened by adverse side effects (Tamminga et al., 2005; Lublin et al., 2005). There are, however, considerable contributions from genetic factors but most likely in a complex interplay between hereditary and environmental components. The risk of developing schizophrenia is considerably elevated among first-degree relatives of persons with schizophrenia, with heritability estimates around 80% (McGuffin et al., 1995; Riley et al., 2006; Sullivan et al., 2012).

1.1.2 Bipolar disorder

Bipolar disorder, also referred to as bipolar affective disorder (and earlier manic-depressive disorder), is a life lasting disorder of slightly higher prevalence than schizophrenia. The current lifetime prevalence estimate is about 0.7% (Sullivan et al., 2012) and both point and 6 or 12 month prevalences are at the same level (see Ferrari et al., 2011). In the GBD 2010 study, three bipolar disorder health states were used: depressive, manic, and residual (Ferrari et al., 2012). By a literature review Ferrari et al. (2012) found that 27% of individuals with bipolar disorder were in a depressive state, 23% were in the manic state, and the remaining 50% were in other states (residual). According to the World Health Organization (1993), 10th revision of the International Classification of Diseases (ICD-10), the disorder is characterized by repeated (two or more) episodes with significantly disturbed mood, both episodes with elevation of mood and increased energy and activity (mania or hypomania) and episodes of lowering mood and decreased energy and activity (depression). The residual states in GBD 2010 involve episodes below the thresholds for depressive or manic state. In terms of YLD, bipolar is number 18 in the currently available tables from WHO[2] and accounts for approximately the same amount (1.8% worldwide) as schizophrenia.

In the ATPD study by Castagnini et al. (2013), the overall incidence rate of bipolar affective disorder was 6.4 per 100,000 person-years and opposite to schizophrenia (see subsection 1.1.1) with an increase in age of onset going from 4.6 for subject 15–24 years of age at onset to 7.3 for the 55–64 years age-band. Also in contrast to schizophrenia, the incidence of bipolar disorder tended to be higher for females (overall female to male IRR=1.1). The median age of onset for bipolar disorder was 42.0 years and again in contrary to schizophrenia, the males had a later onset (median 43.5 years) than females (41.0 years).

Chromosome 12q24.3 and the Slynar locus—a candidate region

Earlier linkage and association studies implicated chromosome 12q24 as a candidate region for bipolar disorder (Ewald et al., 1998; Degn et al., 2001; Ewald et al., 2002) in Danish and Faroese populations. Specifically the microsatellite marker D12S1639 in 12q24.3 was found to have a significant LOD score (base-10 logarithm of odds). Kalsi et al. (2006) fine mapped the 2 Mb (1 Mb = 1e6 base pairs (bp)) region using samples from Denmark and England and found the association signal to be located in a 300 kb (1 kb = 1,000 bp) region surrounding the microsatellite marker D12S307 within what was then called the *Slynar* (*AY070435*) gene. We refer to this as the *Slynar locus* as no coding genes have yet been found, and most transcripts seem to be noncoding. In the NCBI[3] *GenBank* sequence database (Benson et al., 2014) hypothetical protein mRNA *AY070435* on chr12:125,776,769–125,795,825 is overlapping the NCBI Reference Sequence (*RefSeq*) database (Pruitt et al., 2014) gene *LINC00943* (*long intergenic non-protein coding RNA 943*) on chr12:125,787,506–125,796,753. Here the chromosomal positions were obtained from the UCSC[4] Genome Browser (Kent et al., 2002) NCBI March 2006 build, i.e. UCSC human genome (hg) assembly version 18 (hg18) alias the release named NCBI Build 36.1 (NCBI36). For a recent overview of the database resources at NCBI including *dbSNP*[5] (Sherry et al., 2001), we refer to NCBI Resource Coordinators (2014).

Even more recently *insulin-like growth factor 1* (*IGF1*) on 12q23.2 was found to be associated with bipolar disorder in British samples (Pereira et al., 2011)—a gene which is within the region investigated in the Danish study by Ewald et al. (1998).

[3] National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD, USA. `http://www.ncbi.nlm.nih.gov/`

[4] University of California, Santa Cruz, USA. `http://www.ucsc.edu/`

[5] dbSNP: NCBI database of single nucleotide polymorphisms (SNPs) and short insertion/deletion variants (INDELs). `http://www.ncbi.nlm.nih.gov/SNP/`

1.1.3 Panic disorder

Panic disorder is an anxiety disorder with recurrent attacks which are not predictable in terms of specific situations or circumstances and without objective danger. Though attacks typically last only for minutes, the fear of having another attack may cause the patient to avoid specific situations. In cases where other phobias are present, these should be used as main diagnosis rather than panic disorder (c.f. World Health Organization, 1993).

Panic disorder is a quite common disorder with a lifetime prevalence around 4% and a moderate heritability estimate of 48% (Schumacher et al., 2011). The aetiology is not known but high comorbidity with other psychiatric disorders has been found, including bipolar disorder and schizophrenia (see Box 1 in Schumacher et al., 2011).

Most genetic studies on panic disorder (and anxiety disorders in general) have focused on candidate genes, and only two GWAS had been conducted until around 2011 (Schumacher et al., 2011): a German study (Erhardt et al., 2011) and a Japanese (Otowa et al., 2009). The main finding from the German study, the involvement of the *transmembrane-protein-132D* (*TMEM132D*) gene, has been replicated by the Panic Disorder International Consortium (PanIC) (Erhardt et al., 2012).

1.1.4 Shared genetic aetiology

Though schizophrenia is a psychotic disorder, and bipolar disorder is a mood disorder, symptomatically they are very similar, and overlap in genetic aetiology is empirically evident—most recently and convincingly provided by the Cross-Disorder Group of the Psychiatric Genomics Consortium (2013a) of the Psychiatric Genomics Consortium (PGC) estimating a high genetic correlation ($r_{gSNP} = 0.68$) between the two disorders by use of large GWAS samples of 6–12,000 subjects. Here a positive genetic correlation means that cases of one disorder show higher genetic similarity to the cases of the other disorder than to their own controls. The Cross-Disorder Group of the Psychiatric Genomics Consortium (2013a) also found a moderate positive correlation between schizophrenia and major depressive disorder ($r_{g\,\mathrm{SNP}} = 0.43$) and between bipolar disorder and major depressive disorder ($r_{g\,\mathrm{SNP}} = 0.47$). These results are consistent with polygenic scores (International Schizophrenia Consortium et al., 2009) from the recent PGC Cross-Disorder meta-analysis (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013b). In addition, it is known that panic disorder comorbidity exists with both bipolar disorder (Simon et al., 2004; Young et al., 2013) and schizophrenia (Young et al., 2013). In a multinomial analysis of five psychiatric disorders (schizophrenia, bipolar disorder, autism spectrum disorder, attention deficit-hyperactivity disorder, and major depressive disorder), the PGC study identified four statistically significant ($p<5e\text{-}8$) loci: regions on chromosome 3p21 (nearest gene *ITIH3*), chromosome 10q24 (nearest gene *AS3MT*), and single nucleotide polymorphisms (SNPs) from the calcium channel genes *CACNA1C* and *CACNB2*. Another example is the chromosomal region 12q24 (including the Slynar region, see subsection 1.1.2) which has been associated with both bipolar disorder and schizophrenia, see paper 5 (Buttenschøn et al., 2010).

1.1.5 Suicidal behaviour

Suicidal behaviour is a complex phenotype that includes suicidal ideation, attempted and completed suicide (Willour et al., 2012; Pandey, 2013). Co-morbidity with other psychiatric disorders is very often seen with estimates as high as 90% of suicides being concurrent with psychiatric disorders including substance abuse (Arsenault-Lapierre et al., 2004) and in particular mood disorders (Gonda et al., 2012; Pandey, 2013). The co-occurrence with schizophrenia is also notable and Saha et al. (2007) estimated the median of SMRs to be as high as 12.9 for death by suicide for individuals with schizophrenia. In a large Danish population study, Qin (2011) found the risk of suicide to be significantly increased for individuals with hospitalised mental illness. This study also showed that the diagnosis most commonly associated with suicide depends on age and gender: schizophrenia for younger subjects ($\leq$35 years), mood disorders for older individuals ($>$60 years) and alcohol abuse for middle-aged men (35–60 years) (Qin, 2011). Also the clear and well established fact that more males than females commit suicide depends on age in this study, with the male to female ratio decreasing from 3.1 over 1.8 to 1.4 when considering subjects $\leq$35, 35–60 and $>$60 years.

In addition to the high risk factor from being mentally ill, a separate genetic contribution has been found both in association studies (Willour et al., 2012), family studies (Brent et al., 1996; Turecki, 2001), twin studies (Voracek et al., 2007) as well as in a recent Danish adoption study by Petersen et al. (2013). The estimated heritability is approximately between 30 and 55% (Voracek et al., 2007; Willour et al., 2012). Pandey (2013) recently reviewed suicide studies concerning involvement of the serotonergic and noradrenergic systems, neurotrophic factors, and the hypothalamic-pituitary-adrenal (HPA) axis. Abnormalities in the functioning of the serotonergic system have been associated with impulsive aggressive behaviour (Pandey, 2013), and genes encoding proteins involved in the regulation of serotonergic neurotransmission have been investigated in numerous association studies (Tsai et al., 2011; Willour et al., 2012).

1.1.6 Multifactorial aetiologic architecture of psychiatric disorders

Though evidence is building up especially from the large consortium initiatives, much of the genetic basis, also referred to as the genetic architecture, underlying psychiatric disorders is still unknown. With the high heritability estimates usually seen for psychiatric disorders (see e.g. Table 1 in Sullivan et al. (2012)), we should expect that large parts of the aetiologies for psychiatric disorders hide in genetic variation. Nevertheless, the identified sites until recently only accounted for a very small proportion of the variation and the concepts of missing (or hidden) heritability have been subject of much debate and research (McCarthy et al., 2008; Zaitlen et al., 2012). Why and how is it hidden and maybe more importantly how can it be exposed? Is the vulnerability a result of the cumulative polygenic effect of many common genetic variants, the so-called common disease, common variants hypothesis (CDCV)? Or is the overall disease prevalence rather a result of many rare (private) mutations each of which have a large effect, the Multiple Rare Variants hypothesis (MRV) also known as the rare variants hypothesis (CDRV)?

Though boundaries vary, single nucleotide variants are usually categorised as common variants (or synonymously polymorphisms, i.e. SNPs) when the minor allele frequency (MAF) is at least 1% (sometimes higher) and rare variants (usually abbreviated SNVs) when MAF$<$1%, see e.g. Frazer et al. (2009). Very rare SNVs (MAF$<$0.1%) are sometimes referred to as novel or *de novo* SNVs. Following the GWAS era that we have witnessed the last 5–8 years, one of the hottest topics right now is next generation sequencing (NGS)—targeted, exomic or whole-genomic—and not less combinations of sequencing, GWAS and insilico genotyping (imputation). In addition, variation is also attributed to the so-called structural variants which are broadly defined as genetic variants that are not single nucleotide variants (Frazer et al., 2009). This class includes insertion/deletion variants (INDELs), block substitutions, inversions, translocations, and copy number variants (CNVs) which are smaller or larger segments (but $>$ 1 kb) of DNA which have been duplicated (gains) or deleted (losses). As with SNVs the more common CNVs (MAF$>$1%) are referred to as copy number polymorphisms (CNPs). Due to the interest in CNVs, SNPs that are able to tag (capture the variation) these loci have been added to GWAS chips which are otherwise mainly targeting common variants.

We hypothesize a multifactorial aetiologic architecture of psychiatric disorders where the truth is neither CDCV nor MRV but rather a combination in concert with environmental effects, gene-environment interactions (G×E), epigenetic factors, and gene-gene interactions (G×G), i.e. either epistatic effects or statistical deviations from additivity. Moreover, instead of classifying cases into either a single-mutation mechanism under the MVR or polygenic mechanism under the CDCV, Mitchell et al. (2011) argue that a mixed model involving interactions between disease-causing and disease-modifying variants is a biologically more plausible model in schizophrenia. In this mixed model, the polygenic effect does not produce the phenotype itself but instead modifies the highly penetrant mutations. Thus, Mitchell et al. (2011) propose that probably all cases of schizophrenia are dependent on the presence of highly penetrant mutations.

Concerning heritability, which overall measures the proportion of phenotypic variation explained by genetics, this concept is usually divided into broad and narrow sense heritability. Broad sense heritability (H^2) includes both epistatic (G×G), dominance and additive effects whereas narrow sense heritability (h^2) only includes additive effects. The remaining part of the phenotypic variance is usually attributed to environmental variance. Estimates of heritability from GWAS generally ignore dominance and epistatic effects, i.e. they are narrow sense heritability estimates (Zaitlen et al., 2012). Estimates of total heritability usually also ignores epistatic effects, and this may have inflated heritability estimates and thus means that the proportion explained by the additive effects may be correspondingly under-estimated (see Zuk et al., 2012). Yet another measure of heritability is the SNP heritability, i.e. the proportion of the variation in liability to a disease explained by SNPs (Lee et al., 2011). Using this measure Lee et al. (2012) estimated the SNP heritability in schizophrenia to be 23%. Ripke et al. (2013) found this measure to be 32% and explaining as much as 50% of the heritability and thus concludes that disease causing variants tagged by common SNPs may have a crucial contribution to the risk of schizophrenia.

In favour of the MRV hypothesis in schizophrenia are recent results from exome sequencing of case-control (Purcell et al., 2014) and trio (Fromer et al., 2014) samples. The study by Purcell et al. (2014) tested for enrichment of rare alleles in approximately 2,500 genes implicated by large-scale whole-genome studies including both GWAS (common SNPs), CNV and *de novo* SNV studies. Purcell et al. (2014) found that both common SNPs, rare CNVs and rare ($\text{MAF}<0.1\%$) disruptive mutations (e.g. nonsense mutations[6]) were independently and additively enriched in cases. A polygenic burden attributable to many very rare nonsense mutations distributed across many genes was found, though the contribution to the heritability was an order-of-magnitude higher for GWAS variants than for the rare variants. The study by Fromer et al. (2014) identified *de novo* mutations in schizophrenia for multiple sets of functionally related proteins involved in synaptic mechanisms. These findings were replicated in the case-control study by Purcell et al. (2014). A relevant question may of course be, if these findings in schizophrenia apply in other psychiatric disorders too? A Swedish study by Bergen et al. (2012) indicates that this may not be the case—at least not with respect to (larger) CNVs which were found to be enriched in patients with schizophrenia but not in patients with bipolar disorder. On the other hand SNVs were more frequent in both patient groups compared to controls.

[6] Point mutations resulting in a premature stop codon.

Immune system implications in psychiatric disorders

Though only few studies have considered the involvement of immune defects and inflammation in suicide, Pandey (2013) notes that this may be of importance and should be studied further. The possible contribution from infections, inflammations and autoimmune disease to the aetiology of psychiatric disorders has otherwise been investigated and speculated for ages (Yolken et al., 1995). Particularly the involvement of the immune system for psychosis in general and schizophrenia in specific is supported by numerous studies (e.g. Buka et al., 2001; Eaton et al., 2006; Yolken et al., 2008; Xiao et al., 2009; Havik et al., 2011; Benros et al., 2011; Benros et al., 2012; Fillman et al., 2013). Kirkpatrick et al. (2013) give an update of key findings on inflammation in schizophrenia and give a longer list to guide future research. Association of inflammatory state, autoimmune processes and infections has also been suspected with bipolar disorder (Eaton et al., 2010; Leboyer et al., 2012) and other mood disorders (Benros et al., 2013), panic disorder (Salazar et al., 2012) and anxiety disorder Chen et al. (2013). Furthermore, the many studies implying the major histocompatibility complex (MHC) region on chromosome 6 are consistent with a possible connection between the (auto)immune system and mental disorders (e.g. Stefansson et al., 2009; International Schizophrenia Consortium et al., 2009; Shi et al., 2009; Havik et al., 2011; Ripke et al., 2013). Involvement of pathways in the immune systems is also prevailing in other diseases of the brain, e.g. in neurodegenerative diseases (Ramanan et al., 2013) such as Parkinson's disease (Holmans et al., 2013) and Alzheimer's disease (Lambert et al., 2010). One of the pathways involved in the immune system is the lectin pathway of complement activation. Two key components for this activation process are mannan-binding lectin (MBL) and MBL-associated serine protease-2 (MASP-2) (Garred et al., 2009). MBL deficiency is the most common hereditary defect in the human immune system (Thiel et al., 2006) and is known to be associated with the presence of three nonsynonymous mutations in exon 1 of the gene *MBL2* encoding the protein (Garred et al., 2006). Another three polymorphisms from the promoter region of *MBL2* explain much of the remaining variation in MBL serum concentration, and six common haplotypes formed by these six variants correlate with different levels of MBL (Garred et al., 2006), and the deficiency is very heterogeneous (Heitzeneder et al., 2012; Mayilyan, 2012). Moreover, not just deficiency but also increased activity of the complement pathway has been observed in patients with schizophrenia (Mayilyan et al., 2008), mainly in MASP complexes. It is also worth mentioning that some infections may play a role in the development of autoimmunity (Galli et al., 2012), and that inflammation is inherent to both states.

Other non-genetic factors for psychiatric disorders

Many non-genetic factors have been speculated as potentially predisposing (risk factors) for psychiatric disorders. Here we will mention some but the list is far from being complete. For a recent overview in schizophrenia, we refer to Torrey et al. (2012). Various risk factors involving the immune system have been suggested: in maternal infections during pregnancy with Toxoplasma gondii (Brown et al., 2005; Mortensen et al., 2007), cytomegalovirus (Borglum et al., 2013) or herpes simplex virus 2 (Mortensen et al., 2010). Also more general vulnerability for severe (hospital-treated) infections has been associated with schizophrenia (Nielsen et al., 2013a; Nielsen et al., 2013b). Other factors concerning embryonic stage and birth include obstetrical complications (Nicodemus et al., 2008) and small for gestational age (relatively low birth weight) (Nielsen et al., 2013c). In Larsen et al. (2010) we found prematurity and low birth weight to be risk factors for subsequent development of affective disorder (especially depression) and schizophrenia. Parental age has also been found to influence the risk of psychiatric disorders. Though varying in a complex fashion with increased risk for some disorders but little or no effect for other, McGrath et al. (2014) concluded that offspring of younger mothers (less than 25 years of age) and older fathers (older than 40 years) are at higher risk for various mental health disorders. Somewhat mysteriously, the association between schizophrenia and paternal age has found to relate to the fathers age at birth of his first child rather than at conception of later children (Petersen et al., 2011). Seasonality of birth has long time been mentioned as a risk factor (Torrey et al., 2012) but is probably not that important. Nevertheless, in Sorensen et al. (2013) we found an association between being born in the autumn and risk of clozapine treatment, which may be seen as a measure of treatment resistance and severity of illness. The patterns were not very clear though. Finally, cannabis use has been implicated in schizophrenia (Torrey et al., 2012), and in Arendt et al. (2005) we found cannabis-induced psychotic symptoms to be an important indicator for subsequent development of severe psychopathological disorder.

Gene-environment (G×E) interaction

The involvement of non-genetic factors (often stated as environmental factors) may also be less direct as a moderator of genetic risk factors via interaction. Such gene-environment interactions (G×E) may be an important source of complexity to the aetiology of psychiatric disorders. Nevertheless, only few findings have been reported, possibly due to at least two complicating factors: the need for large samples in concert with information on the environmental exposure. The comprehensive Danish study *i*PSYCH will include such information, drawing partly on the Danish registers and partly on the ability to extract information from neonatal dried blood spots about exposures to the fetus. The significant interaction between genetic markers and maternal infection with cytomegalovirus identified by Borglum et al. (2013) exemplifies how the immune system may also be involved via G×E interactions.

The number of methods and software available for G×E analysis is overwhelming and there is no clear winner or gold standard. A trick that is often used when comparing statistical methods is to simulate data with known properties and investigate how well the methods manage to identify these characteristics. Finding and choosing the most appropriate method and software to simulate genotypic data may also be difficult in itself, but a web site was recently established to accommodate this process (Peng et al., 2013). Moreover, a thorough review of state of the art software for computer simulations of population and evolution genetics can be found in Hoban et al. (2012). A very general forward-time simulator with the ability to simulate individuals with genotypes under many evolutionary scenarios is *simuPOP* (Peng et al., 2005; Peng et al., 2012). The *simuPOP* simulation environment is based on Python[7] and the core software/scripts as well as user contributions can be downloaded freely[8].

1.1.7 Summarising signals

An increasingly important issue in genetic research of today, e.g. in GWAS and even more in NGS, is correction for multiple testing to avoid publishing findings that are merely falsely rejected observations under the null hypothesis, i.e. false positives. Here it is truly important to address the enormous increase of type I errors introduced when performing hundreds of thousands or even millions of statistical tests simultaneously. The commonly accepted genome-wide threshold for single-marker association tests has become 5e-8 but whole-genome searches for interactions between the markers or with other factors like environmental disease predisposing exposures obviously involve drastically many more tests and thus the need for an even lower p-value

[7] `http://www.python.org`

[8] `http://simupop.sourceforge.net/`

threshold to control the risk of false discoveries. As an example, a threshold of 1e-12 was proposed by Becker et al. (2011) when doing all SNP-SNP interactions for SNPs on a 500K chip. Lowering the threshold comes at the unfortunate but inevitable expense of increasing the probability of type II errors and thus lowering the power to detect association. Really, this is a double-edged sword of doing extremely many tests simultaneously—not only are we prone to get more errors, if we correct for this we are faced with a low power to detect effects. Relatively large effects would help (but effects are often not large) or we would need huge sample sizes to compensate this loss of power. The latter is essentially happening via worldwide collaborations in big consortia like the PGC and the former is sometimes the gain from considering G×E effects (e.g. Borglum et al., 2013).

Yet another lane to follow are efforts to reduce the multiplicity by e.g. summarising tests across predefined regions such as candidate genes or elements of functional pathways. Along these lines are also methods for interaction analysis using aggregation or multiple steps to reduce the number of tests such as multifactor dimensionality reduction (MDR) (Ritchie et al., 2001) and two-step procedures like the G×E approach by Murcray et al. (2009). Summarising test statistics to obtain a combined value may result both in fewer tests and diminish dependencies between tests so that standard procedures apply. A classical approach is Fisher's method for meta-analysis (*Fisher's combined probability test*)

$$-2\sum_i \log p_i \sim \chi^2_{2k}, \quad p_1, \ldots, p_k,$$

where the k p-values are assumed to be independent (Fisher, 1932). But caution is needed as the distributional approximation in Fisher's method like Bonferroni methods relies on the independence assumption, and the combined test may be anti-conservative if the assumption is invalid. Furthermore, it can be problematic that the groups are defined beforehand as it may turn out not to be the appropriate grouping to work with—e.g. this might exclude important but non-coding DNA outside the defined regions such as the many regulatory elements recently mapped by the ENCODE Project (ENCODE Project Consortium et al., 2012).

The idea of building up evidence by aggregation of signal over a segment may be motivated by the following example. Suppose we have calculated p-values of 20 tests ordered by position, $p_1, \ldots, p_{20}$ within some region. Under the null hypothesis of no association, it would be unlikely to observe a longer sequence of single-marker p-values below the level of significance, $\alpha = 0.05$ say. Now let $Z_k = 1$ if $p < 0.05$ and $Z_k = -1$ otherwise and suppose that in a concrete but hypothetical example these evolve as depicted in figure 1.1. The question is then of course how to quantify the signal that appears to build up between markers 2 and 14, and how to take care of e.g. distance between markers. In paper 6 we propose a method that can handle this kind of summation.

1.2 Aims

The main hypothesis of the PhD study was that gene-gene (G×G) and gene-environment (G×E) interactions play an important role in the aetiology of complex diseases like psychiatric disorders. Therefore, the main aim of the study was to investigate methodological aspects of and apply methods from statistical genetics that take interactions into account. In addition to this we considered issues concerning single-marker tests, analysis of sex chromosomes, and use of haplotypes and other means of accumulating signals from multiple markers.

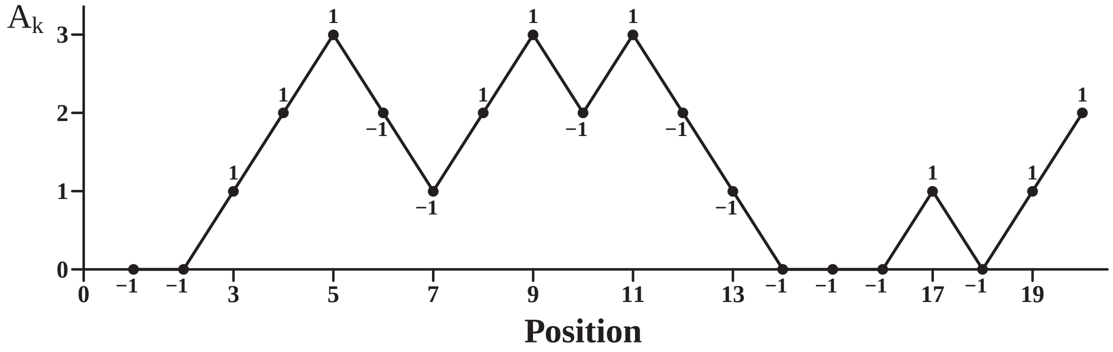


Figure 1.1 Motivating example of aggregating signals. A hypothetical example of building up evidence from a sequence of tests. Z_k, defined as $Z_k = 1$ if $p < 0.05$ and $Z_k = -1$ otherwise, are the values shown above the circles. The x-axis indicates the position in the sequence and the y-axis is the sum of Z_k's with truncation at zero, i.e. the function $A_k = \max\{0, Z_k + A_{k-1}\}$.

The original aims for the included case-control studies (papers 1, 2, 4 and 5) were related to health sciences and thus not fully consistent with the aims of the present PhD study. Some of the underlying statistical issues and aspects will be treated in more depth in chapter 3. Paper 3 and 6 are different as they are mostly based on computer simulated data and aim at developing software, methods and achieving basic knowledge of pros and cons of existing methods. Still, though, with the practical viewpoint that this knowledge is later going to be used for or assist decisions taken in health science studies using real data.

1.2.1 Aims of paper 1 and 2: MBL and MASP-2 study

Papers 1 and 2 were motivated by the hypothesis that defects of the immune system may increase risk of psychiatric disorders, see subsection 1.1.6. The main aim of paper 1 was to investigate if schizophrenia is associated with MBL and MASP-2 serum concentration and/or with genetic variants of the genes coding these proteins: *MBL2* on chromosome 10q21.1 and *MASP2* on chromosome 1p36.22. We also explored disease association with protein levels after adjustment for the genetic polymorphisms as well as with inclusion of G×G interactions. In paper 2 the same objectives were applied to bipolar disorder and panic disorder.

A specific aim in relation to MBL serum concentration was to figure out how to model quantitative traits that are censored due to detection limits of the measurement method, as this was the case for MBL protein in a relatively large proportion of the subjects.

Originally we planned to analyse the three disorders simultaneously (Foldager et al., 2009b) but it turned out to be difficult to write the results in just one paper. In the Results chapter, we will present some of these combined analyses. If for nothing else, a means of making comparison of the two papers and three disorders easier.

1.2.2 Aims of paper 3: G×E simulation study

To guide decision of which G×E methods to use, we are doing a simulation study to compare some of the most promising or frequently used methods, from multi-step regression analyses to machine learning. The intention being to characterize performance of a number of G×E methods in a wide range of standardized scenarios to facilitate informed choices in future and ongoing projects such as *i*PSYCH. We intend to consider a range of scenarios by varying minor allele frequencies, sample sizes, effect sizes and penetrance models, and to compare methods of the following kind: two-step analysis, MDR (Ritchie et al., 2001), logic regression (Ruczinski et al., 2003), random forests (Breiman, 2001), artificial neural networks, genetic programmed neural networks.

Paper 3 presents the first steps of this simulation study. The foremost aim of this part of the study was to decide how to simulate individual-based genotypic data including G×E and maybe G×G interactions. It turned out to be necessary to revise and extend the method/software to meet our requirements. Secondly, after simulating data for a few scenarios and checking that data generated conform to the models intended, a few selected methods were to be analysed to shed light on how many methods and scenarios it is reasonable (with respect to time) to compare. We touch superficially on how to fairly compare the methods.

1.2.3 Aims of paper 4: Suicide study

The aim of the study on completed suicides sent to autopsy in Denmark, covering the years 2000–2007, was to investigate the association between suicidal behaviour and candidate genes in the serotonergic system (see subsection 1.1.5). Five genetic markers were chosen on the basis of findings from other populations, see subsection 2.2.3.

One of the markers was from the X chromosome and this induced the need to investigate how to analyse such a marker, since analysis of sex chromosomes likely requires different approaches. A biological question influencing this choice of statistical model is whether or not an assumption of inactivation of one of the female X chromosomes early in development is plausible. Another methodological issue in paper 4 stems from the fact that the gender distributions in cases and controls differ substantially. Moreover, we included analyses of the interaction between genetic markers and age and gender.

1.2.4 Aims of paper 5: Slynar study

Paper 5 is an example of a focused approach to narrow down a previously found signal to search for more precise positions of disease causing mutations and functional implications. The purpose of the study in paper 5 was to reexamine the association to the Slynar locus on chromosome 12q24.3 (see subsection 1.1.2) in Danish patients with bipolar disorder and in Danish patients with schizophrenia. The aims were to replicate earlier findings in an independent sample, fine map the Slynar locus by a selection of SNPs, and to investigate whether the locus is a common susceptibility locus for these two psychiatric disorders. The most significantly associated marker was also genotyped and analysed in a Scottish replication sample. Furthermore, a meta-analysis of this marker was carried out across the Danish and Scottish samples in addition to the British (UK) sample used in Kalsi et al. (2006).

Some special issues in this study were due to the combination of both SNPs and microsatellite markers. Microsatellites are usually much more variable and displaying more than the two alleles normally seen in SNPs. Thus permutation-based methods were used to assess association with these markers, and calculation of linkage disequilibrium (LD) and testing for Hardy-Weinberg equilibrium (HWE) needed to be handled differently. Moreover, we used haplotype analysis to aggregate signal over multiple markers.

We also set out to identify the most abundant Slynar transcripts both in human brain and other tissues, and to identify possible novel transcripts. This part was undertaken by a number of the co-authors (HNB, IMLO, MN, MMH, PK and KVC). Especially I.M.L. Olsen and M.M. Hansen did most of the lab work in Aarhus concerning identification of transcripts while our collaborators at Lundbeck A/S, P. Kallunki and K.V. Christensen, performed expression analyses by quantitative real-time PCR. This part of the study should not be seen as a part of the present dissertation and no further details will be given here but can be found in Buttenschøn et al. (2010) (paper 5).

1.2.5 Aims of paper 6: Landscape method

In stark contrast to such a focused approach as made in paper 5, stand studies considering genetic markers spread over the whole genome, i.e. GWAS. Here it is truly important to address the enormous increase in type I error introduced by performing hundreds of thousands of simultaneous tests. At the same time, attempts should be made to develop or use methods that can utilise the data as much as possible and reveal significant effects or indications even after correction for multiple comparisons which will always be at the expense of a lower power (see also subsection 1.1.7).

In paper 6 we propose a method called *Landscape* that summarises a series of sequentially ordered test values without the need of more or less arbitrary prior grouping. The procedure was inspired by Random Walk theory (Karlin et al., 1990; Karlin et al., 1992) and searches the sequence for a stretch of consecutive values that combined may show evidence of association with that *segment*. The method is general and may be used in other contexts than genetics and genomics, e.g. time-series analysis. The sequentially ordered test statistics, or even more generally any set of sequentially ordered stochastic variables, may be dependent or independent. If they are independent, an approximate distribution of the aggregated value can be devised whereas for dependent variables, bootstrap methods are used to approximate the distribution. We also suggest how to adjust for the multiple testing which is still relevant but at a lower dimension, i.e. accounting for fewer simultaneous tests.

Chapter 2

Materials and methods

2.1 Samples

The samples are more thoroughly described in the corresponding papers but will briefly be outlined here, divided on phenotype rather than study and including information about sharing across studies. The samples simulated for paper 3 is also sketched. In section 2.1.7, we give a short description of a data set which was used as a proof of concept example for the *Landscape* method presented in paper 6.

2.1.1 Schizophrenia

Two hundred and four Danish patients with schizophrenia, 94 (46%) females and 110 (54%) males of Danish Caucasian descent at least three generations back were used for the study presented in paper 5 (Buttenschøn et al., 2010). The patients were interviewed with the semi-structured diagnostic interview Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (Wing et al., 1998), and final best-estimate life-time diagnoses were achieved by consensus of two experienced psychiatrists. The patients fulfilled the DSM-IV (American Psychiatric Association, 1994) and ICD-10 (World Health Organization, 1993) diagnostic criteria for schizophrenia (ICD-10 F20).

A subset consisting of 100 of these patients, 50 of each gender, were also used in paper 1 (Foldager et al., 2012).

2.1.2 Bipolar disorder

The Danish bipolar samples used in paper 5 (Buttenschøn et al., 2010) consisted of 166 patients with bipolar disorder assessed by the same means as the schizophrenia sample described in subsection 2.1.1. The patients with bipolar disorder fulfilled the ICD-10 criteria for bipolar affective disorder (ICD-10 F31) and the DSM-IV criteria for bipolar I disorder. The sample was composed of 84 patients (42 females and 42 males) from the same study as the schizophrenia patients and 82 patients from another Danish study; 48 females (59%) and 34 males (41%). One male individual of these 82 patients was excluded due to large degree (>50%) of failing genotyping (missing genotype values).

The group of 84 patients supplemented with 16 individuals from the other group was selected for the study presented in paper 2 (Foldager et al., 2014). The additional 16 patients (6 females and 10 males) were simply chosen as the first 16 subjects sorted by identification number. Unfortunately, no material for serum determination was available for these 16 patients (and neither for the other 66 from that study). Thus, final gender distribution (female/male) of the bipolar sample in paper 2 was 50–50 for the serum analyses and 48–52 for the genotypic analyses.

In paper 5 we furthermore used a replication sample consisting of 162 Scottish patients with bipolar disorder. We miss gender information for a few Scottish individuals, but from what we have, it appears to be fifty-fifty females and males. The diagnoses of the Scottish cases were determined by case notes reviews and personal interviews using the Schedule for Affective Disorders and Schizophrenia – Lifetime version (SADS-L) (Endicott et al., 1978). The final diagnoses according to the DSM-IV criteria were reached by consensus between two experienced psychiatrists (c.f. Severinsen et al., 2006). Genotyping failed for 14 of the Scottish patients, and they were therefore excluded from the analyses.

2.1.3 Panic disorder

A total of 100 patients with panic disorder without a history of bipolar disorder obtained from previous genetic studies were used for paper 2 (Foldager et al., 2014). The patients, 79 females and 21 males, had been diagnosed with the SCAN interview (Wing et al., 1998) and fulfilled a life-time, best-estimate diagnosis according to ICD-10 (F41) and DSM-IV. To minimise the effect of population stratification, recruitment was restricted to individuals of Danish ancestry for three generations.

2.1.4 Suicidal behaviour

The nation-wide Danish association study of completed suicide by Buttenschøn et al. (2013) (paper 4) considered all deaths sent for autopsy by the police between 2000 and 2007 to confirm suspected suicide. Suicide cases were classified as violent or non-violent (death by poisoning) and were obtained post-mortem from the three Danish forensic centres in Aarhus, Odense and Copenhagen. The biological material from cases used in this study was muscle tissue samples collected at autopsy. Psychiatric registrations, gender, date and place of birth, citizenship and place of residence at death, as well as place of birth of their parents were obtained from the Danish Psychiatric Central Register (Mors et al., 2011) and the Danish Civil Registration System (Pedersen et al., 2006). More than half (57%) of the cases also had a record in the Danish Psychiatric Central Register (Mors et al., 2011): 10% with a schizophrenia spectrum disorder diagnosis (ICD-10 F21-F29, F60.0 and F60.1), 10% with affective disorders (F30-F39), 14% with substance dependence (F10-F19) and 23% other diagnoses. Further details and gender specific proportions can be found in Table 2 of paper 4. To ensure primarily Danish and Caucasian ethnicity, cases born outside Denmark were excluded unless both parents were Danish born. Individuals without a valid personal identification number (CPR number) were also excluded. After exclusions, the sample consisted of 572 suicide cases: 209 (37%) females and 363 (63%) males.

2.1.5 Control samples

Obtaining a good sample of controls, preferably healthy subjects not suffering from the disease or disorder investigated, can be a challenge, and quite often some of the used samples have not been thoroughly screened for the phenotype of interest. Moreover, using the same controls in multiple studies is probably more the rule than the exception—and this is also the case for the studies included in this thesis as we will now describe in some detail.

Controls for MBL and MASP-2 study (paper 1 and 2)

Three hundred and fifty healthy, psychiatrically unscreened Danish volunteer blood donors were obtained as controls for the studies in paper 1 and 2 concerning the involvement of MBL and MASP-2 in psychiatric illness (Foldager et al., 2012; Foldager et al., 2014). Restrictions defined by the ethnical committees preclude information of ethnic origin and other demographics about the controls, but they were expected to be of mainly Western European descent. Since a health questionnaire must be completed and approved before blood donation in Denmark, none of the donors should suffers from a current infectious disease.

Controls for Suicide study (paper 4)

The control samples for the suicide cases were obtained from Danish working and student populations. The controls from the working population (WP controls) were all part of the Danish PRISME study (Kolstad et al., 2011), which recruited workers from work units within several large public workplaces in Aarhus, Denmark. The WP controls were screened for depression and recent suicidal thoughts by questionnaire. WP controls born outside Denmark, with a record in the psychiatric register or without a valid CPR number, were excluded. The student population controls (SP controls) were unscreened medical students recruited as controls for other genetic studies. For these we were unable to access personal data except for gender and ethnicity. At inclusion, they confirmed that both parents and all four grandparents were born in Denmark, and since they were sampled during their first two years as students, we assumed that they were less than 35 years old. After exclusions, the total number of controls was 1049: 545 questionnaire screened WP controls and 504 unscreened SP controls (see table 2.1).

Table 2.1

Samples for the suicide study (paper 4). Number of subjects in each group and for each gender: cases (completed suicide), working population (WP) controls screened for psychiatric illness, student population (SP) controls consisting of unscreened medical students, and combined (All) controls. The figures in parentheses are column proportions except from the figures in the total row where the proportions are out of the grand total.

Gender	Cases	WP controls	SP controls	All controls	Total sample
Females	209 (0.37)	443 (0.81)	320 (0.63)	763 (0.73)	972 (0.60)
Males	363 (0.63)	102 (0.19)	184 (0.37)	286 (0.27)	649 (0.40)
Total	572 (0.35)	545 (0.34)	504 (0.31)	1049 (0.65)	1621

Originally, the plan was to include only the WP controls. However, as the workplaces and professions of these public workers have a major dominance of females (29% were nurses and 18% social work and counseling professionals, c.f. Kolstad et al. (2011)) and as completed suicide is more prevalent in males (see e.g. the counts in Qin, 2011), we were confronted with a very different gender distribution in cases and controls (see table 2.1). Especially the fact that the number of male controls were less than a third of the number of male cases seemed very unfortunate, not least considering the power to detect differences between male cases and controls. At first we then decided to use the male SP controls although we could not screen them for mental illnesses by use of questionnaire or register information. Nevertheless, the inclusion of the male students only minimised the problem by raising the male proportion in the controls from 19% to 39% in contrast to the 63% males in the case group. Thus, confronted with the fact that we still had to take an uneven distribution of genders into account, we decided to also include the female part of the student population - expecting this to increase the power to detect association, apart from anything else.

Differences in gender distribution between cases and controls is a well known problem when using standard control sets and may, as noted also by Clayton (2009), not just be the result of a bad design. A prominent example in this respect is the common set of controls used for several diseases in the Wellcome Trust Case Control Consortium (WTCCC) study (Wellcome Trust Case Control Consortium, 2007).

Controls for Slynar study (paper 5)

Three hundred and eleven ethnically matched controls of Danish Caucasian descent three generations back were included in the study published in paper 5 (Buttenschøn et al., 2010). Of these 191 is a subset of the SP controls used in the suicide study (see subsection 2.1.5), 115 (60%) females and 76 (40%) males, while the other 120 were healthy controls originally used in a study on breast cancer and therefore all females. Two of the breast cancer controls were excluded due to large degree (>50%) of missing genotypes. All Danish control individuals were unscreened for psychiatric disorders.

The Scottish replication sample included 200 ethnically matched controls from the same population with approximately fifty-fifty females and males. Scottish controls were recruited from Scottish National Blood Transfusion Service donors and screened to exclude people with serious chronic illness and those taking any form of medication apart from contraceptive pill and hormone replacement therapy (c.f. Borglum et al., 2001). Thirteen of the Scottish controls were exclude due to failed genotyping.

2.1.6 Simulated case-control samples (paper 3)

The simulation of case-control samples for paper 3 was a multi-step procedure. From an initial population of unrelated individuals and a selected set of SNPs, a *base population* was generated by expansion of the initial population for a large number of generations. This base population was used to generate offsprings for which affection status was imposed under certain scenarios and case-control samples thereby drawn using a rejection sampling algorithm. One hundred case-control samples of 5,000 affected and 5,000 unaffected individuals were generated for each of the 16 scenarios. Further details are given in section 3.2 and in paper 3.

2.1.7 WTCCC data for *Landscape* (paper 6)

In a combined analysis of two genome-wide association studies (GWAS) by Sklar et al. (2008) signal of association with bipolar disorder was found for the SNP rs1006737 in the *calcium channel, voltage-dependent, L type, alpha 1C subunit* (*CACNA1C*) on chromosome 12p13.33. The signal was found after combining online p-values from the WTCCC bipolar sample (Wellcome Trust Case Control Consortium, 2007) with p-values obtained from a combined sample of bipolar I patients from the Systematic Treatment Enhancement Program for Bipolar

Disorder (STEP-BD) and University College London (UCL). The result was further confirmed by adding a third GWAS (Ferreira et al., 2008), and the SNP was also found to be associated with schizophrenia in a Danish study (Nyegaard et al., 2010), suggesting *CACNA1C* as a common risk gene for both bipolar disorder and schizophrenia. The involvement of *CACNA1C* in schizophrenia was latest confirmed in Ripke et al. (2013).

As a proof of concept of the *Landscape* method, we re-examined the bipolar data from WTCCC (Wellcome Trust Case Control Consortium, 2007) for *CACNA1C* extended by a buffer zone to both ends of the gene region to avoid edge effects. Genotypes for a total of 204 SNPs were used after filtering according to the description in WTCCC (Wellcome Trust Case Control Consortium, 2007). Usual trend test p-values were calculated using logistic regression on the original data and on 999,999 permutation samples obtained by randomly shuffling the original case-control labels and used for calculation of permutation-based p-values in *Landscape*.

2.2 Genes, DNA, genotyping and serum measures

2.2.1 MBL and MASP-2 study (paper 1 and 2)

Genomic DNA was extracted from whole blood to investigate associations with genetic markers located in the genes coding for MBL and MASP-2: *MBL2* located at 10q21.1 and *MASP2* located at 1p36.22. In *MBL2*, three SNPs from the promoter region (rs11003125, rs7096206, rs7095891) and three nonsynonymous mutations of exon 1 (rs5030737, rs1800450, rs1800451) were genotyped. Figure 1 of paper 1 (Foldager et al., 2012) shows the positions of the markers in *MBL2*, see also figure 4.1 in subsection 4.1. In *MASP2*, one point mutation rs72550870 (also referred to as D120G) was genotyped.

In *MBL2*, the alleles for the three promoter SNPs are usually designated H/L, Y/X and P/Q, and the three variants of exon 1 are referred to as D, B and C, while the wild type variant is denoted A. Due to linkage disequilibrium, only seven haplotypes (HYPA, LYQA, LYPA, LXPA, LYPB, LYQC and HYPD) are commonly found from these markers, with HYPA being the most frequent in samples of European ancestry. However, an additional haplotype LYPD was found in a single control individual probably originating from a recent intragenic recombination event between HYPD and LYPA or LYPB (Boldt et al., 2010). This subject was excluded to avoid dealing with this extra haplotype. The variants in exon 1 are sometimes combined into a diallelic locus with alleles A and O (either of D, B and C) and then combined with the Y/X marker to form two-marker haplotypes YO, YA and XA (XO never observed). The multilocus genotypes obtained from this two-marker haplotype can be classified by their associated level of MBL in serum: high (YA/YA, YA/XA), intermediate (XA/XA, YA/YO) or low (XA/YO, YO/YO) (Olesen et al., 2006). Another reason for using this classification is the induced reduction of categories; 7 haplotypes imply 28 possible multilocus genotypes (of which we see 26—LYPA/LYQC and LYQC/LYQC were not observed), and using a factor with that many categories in relatively small samples inevitably induces cells with low counts which may cause numerical problems in the analyses. The main reason for including the grouping in papers 1 and 2 was, however, to ease comparison to earlier studies and we otherwise recommend using the more detailed genotype grouping (c.f. Foldager et al., 2012).

The concentration of MBL and MASP-2 in serum was measured. For MBL, a detection limit of 10 ng MBL/ml serum applies. Classification of MBL deficiency is not fully solved (Dommett et al., 2006), and various serum levels have been suggested: <10, <50, <100, <500 ng/ml. We referred to the following MBL levels as: low/deficient: <100, intermediate: 100–400, normal:

>400 ng/ml. These levels were chosen to resemble the guiding levels given by the State Serum Institute (SSI)[9] at that time (August 2006). Consistent with the uncertainty regarding clinical relevant levels, it can be noted that the current reference ranges from SSI are: <100, 100–500 and >500 (web page checked February 6, 2014). It should also be noted that the MBL levels associated with the two-marker genotype groups mentioned above were chosen a bit differently in Olesen et al. (2006), where low was <200 and high was >800 ng/ml.

2.2.2 Regions and SNPs for the G×E simulation study (paper 3)

Mimicking a strategy that might be used to reduce the number of markers investigated for interactions, we selected the putative genetic linkage regions for schizophrenia in Caucasians on the basis of a meta-analysis by Ng et al. (2009)[10]: 2q33.3–36.3 (206.3–228.3 Mb), 3p14.1–q13.32 (71.6–120.2 Mb), 5q31.3–35.1 (141.8–167.7 Mb), 6p21.31–12.1 (33.9–56.6 Mb), 8p22-12 (15.7–32.7 Mb), and 16p13.12–q12.2 (13.2–51.5 Mb). From these regions, we chose SNPs that were present on the Illumina HumanHap550 chip and common to all 11 HapMap3 populations (International HapMap 3 Consortium, 2010). We added a buffer zone of 10% of the region size at each end of each region to avoid possible edge effects. To speed up calculations further for inclusion in the present dissertation, we decided to use only the two chromosomal regions harbouring the disease predisposing loci (DPLs).

At first, SNPs were restricted to have a MAF>0.05 in the base population (Foldager et al., 2013). However, to avoid numerical problems we decided to only include SNPs with a genotype frequency of at least 0.05 for the minor allele homozygote. Note that under random mating and thus Hardy-Weinberg proportions, a MAF above 0.05 only corresponds to a genotype frequency above 0.0025. This might be a valid approach for single-marker analyses, but becomes a problem, numerically, for G×G and G×E interaction analyses. A total of 5,500–6,000 SNPs remained for the analyses, see further details in paper 3.

2.2.3 Suicide study (paper 4)

One of the major candidate genes for suicide is the serotonin transporter gene, *solute carrier family 6 member 4* (*SLC6A4* or *5-HTT*) located on chromosome 17q11.2 and involved in the reuptake of serotonin in the synaptic cleft. Especially an INDEL, known as the serotonin transporter linked polymorphic region (5-HTTLPR also referred to as rs4795541), and a nearby SNP rs25531 (both from the 5' promoter region of *SLC6A4*) have been intensively studied in psychiatry, psychology and neuroscience for many disorders (Caspi et al., 2010) including suicidal behaviour (Gonda et al., 2011; Tsai et al., 2011; Willour et al., 2012). These two markers were therefore selected. Earlier studies considered only two alleles of 5-HTTLPR: a short (*S*) variant associated with a lower expression and a long (*L*) variant (Gonda et al., 2011). However, the A to G substitution of the SNP rs25531 has been shown to modulate the effect of the *L*-allele to behave (almost) like the *S*-allele during transcription (Hu et al., 2006) resulting in a lower expression. In the present study, PCR products were directly used for the 5-HTTLPR genotyping to reveal a long (*L*=405 bp) or short (*S*=361 bp) amplicon size. Further procedures on 5μl of this PCR product were carried out to reveal also the rs25531 SNP by the following visible fragments: L_A=340 bp, L_G=166 bp and *S*=297 bp. Using the convention described by Parsey et al. (2006), the combined genotypes of 5-HTTLPR and rs25531 were also reclassified according to

[9] http://www.ssi.dk

[10] http://www.szgene.org/linkage.asp

functional activity: $\{S/S, S/L_G, L_G/L_G\}$ (low expression), $\{S/L_A, L_G/L_A\}$ (medium expression) and $\{L_A/L_A\}$ (high expression). To make notation less cumbersome, we refer to these classes as SS+SL_G+L_GL_G, SL_A+L_GL_A and L_AL_A, respectively.

Also other components of the serotonergic system have been investigated in relation to suicidal behaviour, not least the two *tryptophan hydroxylase* genes *TPH1* and *TPH2* (Tsai et al., 2011; Willour et al., 2012) involved in the initial and rate-limiting step in the synthesis of serotonin. We selected rs1800532 in intron 7 of *TPH1* on chromosome 11p15.1 and rs1386494 in intron 5 of *TPH2* on chromosome 12q21.1 as these two SNPs have been associated with both suicidal behaviour (Li et al., 2006) and completed suicide (Zill et al., 2004).

Another candidate gene for suicide is *monoamine oxidase-A* (*MAOA*) on chromosome Xp11.3, encoding an enzyme involved in degradation of serotonin, noradrenalin, adrenalin and dopamine. The dopaminergic system may be related to impulsivity. We genotyped a functional untranslated variable number tandem repeat (uVNTR) within *MAOA* (*MAOA*uVNTR), with alleles determined according to the number of repeats (2, 3, 3.5, 4 and 5 repeats). These variants affect the transcription of *MAOA* and is often classified according to functional activity (Deckert et al., 1999): 2 and 3 repeats (low expression); 3.5, 4 and 5 repeats (high expression). In addition to these alleles, we identified a new allele corresponding to 4.5 repeats, and we grouped this with the high expression alleles. High expression alleles have been associated with violent suicide attempts in males (Courtet et al., 2005).

DNA from suicide victims was extracted from paraffin-embedded muscle tissue samples and genotyped. Moreover, frozen muscle tissue was available and improved DNA quality for approximately 12% of the cases. DNA from control individuals was extracted from whole blood using standard procedures. Samples with no visible DNA or contaminated with bacterial DNA were excluded; as were samples that looked badly degraded and unlikely to be amplified successfully for the larger fragments (>300 bp) required for genotyping of markers in *SLC6A4*. Further details on DNA extraction and genotyping are given in paper 4 (Buttenschøn et al., 2013).

2.2.4 Slynar study (paper 5)

Genomic DNA was extracted from whole blood from the Danish patients and controls using standard methods and genotyped for 13 microsatellites and 9 SNPs. To enable identification in the figures, we named these 22 markers consecutively according to position: m1, m2, ..., m22. Positions and identification numbers are shown in Figure 1 and Table 1 of paper 5 (Buttenschøn et al., 2010). The selection criteria for the 13 microsatellite loci were based on previous positive findings (Degn et al., 2001; Kalsi et al., 2006), and 5 SNPs were selected based on the positive findings in Kalsi et al. (2006). Additionally, 4 SNPs were chosen on the basis of a functional approach with two being located within exons—rs3803149 (m7) and rs1194029 (m13)—and two located in the proximal promoter region of Slynar transcripts—rs4765449 (m6) and rs1194050 (m11), see paper 5 for a further description. In the Scottish sample we only genotyped the SNP rs7133178, which was the most significantly associated marker with bipolar disorder in the Danish sample.

2.3 Statistical analysis

When reporting parameter estimates (coefficients or odds ratios), we follow the recommendation of Louis et al. (2009). As an example, suppose we want to write an estimated odds ratio (OR) and its 95% confidence interval (CI) consisting of a lower (L95) and an upper (U95) bound, which is often written as: OR (L95–U95). In the notation of Louis et al. (2009) this is written as: ${}_{L95}\,OR\,{}_{U95}$. In some of the tables (see e.g. table 4.2) we have extended this notation by putting the corresponding p-value above the estimated odds ratio: ${}_{L95}\,\overset{P}{OR}\,{}_{U95}$. This saves a lot of space in the tables, enabling us to include for example all three patient groups from the MBL/MASP-2 studies (paper 1 and 2) in the same table and thus facilitating comparison.

Generally, we have chosen a significance level of 5%, and most analyses were carried out using either the commercial statistical software Stata[11], the non-commercial software R[12] (R Core Team, 2013) or a combination of both—exact versions of course varying over time. In addition, various non-commercial commandline-base softwares were used.

2.3.1 MBL and MASP-2 study (paper 1 and 2)

Single-marker genotypic associations were assessed using logistic regression assuming an additive model on the logarithmic (log)[13] scale. The resulting odds ratio indicates the effect of each extra copy of the mutation allele. Hence, the OR between the two homozygote variants is the square of the reported OR. Similarly, the additive effects of carrying 0, 1 or 2 copies for each of the seven haplotypes were considered. Additive effects for each of m multiple SNPs were tested by an m degrees-of-freedom (d.f.) chi-square test that has a corresponding score test which is a generalisation of the Armitage trend test (Balding, 2006).

MBL and MASP-2 in serum were analysed by log-transformed concentrations to account for non-normality. Standard linear regression was used for the analysis of MASP-2, whereas Tobit regression (Amemiya, 1984) was applied for MBL to handle observations below the detection limit by censoring techniques. A categorisation like the low/intermediate/normal classification (see subsection 2.2.1) would also solve this problem, but at the expense of continuity and thus information loss. Logistic regression was used for analyses of MBL deficiency status ($</\geq$100 ng/ml) and MBL serum detection status ($</\geq$10 ng/ml).

2.3.2 G×E simulation study (paper 3)

With the intention to show the reasonableness of doing a more comprehensive study of a larger set of G×E methods, we considered a version of the popular machine learning and data mining method MDR (Ritchie et al., 2001), model-based MDR (MB-MDR) (Calle et al., 2008), and one version of the machine learning method logic regression (Ruczinski et al., 2003), logic feature selection (logicFS) (Schwender et al., 2011b). We furthermore intend to compare these with a traditional two-step logistic regression, consisting of choosing a subset of SNPs followed by an exhaustive search for significant interactions with an environmental variable. Details on MB-MDR and logicFS are given in section 3.6 and in paper 3.

To check for main effects and two-way interactions, we also applied the *boosted one-step statistics* (BOSS) method by Voorman et al. (2012) to search for G×E interactions, and the *boolean operation-based screening and testing* (BOOST) method by Wan et al. (2010) to examine all pairwise SNP-SNP interactions. Moreover, the BOOST software outputs single-marker test statistics.

[11] Stata Statistical Software, College Station, TX: StataCorp LP. http://www.stata.com

[12] R: A Language and Environment for Statistical Computing. http://www.r-project.org

[13] If not specified explicitly as base-10 (log10) we think of the natural logarithm (base e).

2.3.3 Suicide study (paper 4)

The association between genetic markers and the case/control phenotype was investigated using logistic regression. In order not to have age as a continuous variable (not available for all controls), we grouped age in three groups: <35, 35–49, ≥ 50 years. The results are presented from the relevant conditional logistic regression model when no other significant interaction than between gender and age-group exist, corresponding to stratification on gender and age-group.

Gender, age-group and their interaction were associated with phenotype by sampling (see table 2.1), and these factors were therefore included as covariates in the regressions by default. Clayton (2008) and Clayton (2009) note that to gain (or retain) power, stratification by gender should be avoided if the null hypothesis of no association between gender and genotype or allele frequencies can be assumed. Furthermore, under certain conditions, efficiency would be gained by excluding such factors even when heavily associated with phenotype (Clayton, 2008). Supposedly the same arguments hold for the age group variable. Nevertheless, we decided to retain them, as we believe they might be proxies for other unobserved factors, and furthermore we wanted to test their interaction with genotype. These interaction analyses were not planned a priori.

Genotypes of two-allelic markers were coded by an additive term A_i which is 0, 1 or 2, corresponding to the number of minor alleles that individual i carries, and by a dominance term D_i which is 1 for heterozygote carriers and zero otherwise. This ensures independent tests of specific assumptions of genetic models as recommended by Joo et al. (2009) and Zheng et al. (2009). The reduction to the additive model was tested by testing the null hypothesis that the dominance parameter equals zero. This simpler model, however, was only used when the null hypothesis was clearly not rejected. Details concerning the analysis of the X chromosomal marker *MAOA*uVNTR are given in subsection 3.1.4.

2.3.4 Slynar study (paper 5)

Single-marker genotype-wise and allele-wise Fisher's exact association tests were performed. Permutation-based p-values using 1e6 simulations were used for microsatellites as these are highly variable. Logistic regression was applied on significantly associated markers to determine disease risk in the best fitting genetic model. The models considered, apart from the saturated and the null, were the dominant, additive and recessive models. Odds ratios with 95% CI were estimated. Single-marker tests were supplemented by a sliding window haplotype analysis (Schaid et al., 2002). Linkage disequilibrium in terms of r^2 was estimated for all pairs of SNP markers. LD between microsatellite marker m8 and SNP marker m9 were calculated using the Multiallelic Interallelic Disequilibrium Analysis Software (MIDAS) (Gaunt et al., 2006). Using MIDAS, LD was calculated for each allelic combination between all pairwise combinations of any type of loci. Meta-analysis of m9 across the Danish, Scottish and UK samples was carried out by a stratified logistic regression analysis. Genotypes for the m9 marker (rs7133178) in the UK samples were generated on the basis of the allele frequencies reported in Kalsi et al. (2006) assuming HWE.

Chapter 3

Statistical methods

3.1 Association analysis

The general goal of association analysis is to find significant differences between affected and unaffected individuals in polymorphism on loci that might then have a role in increasing (risk alleles) or decreasing (protective alleles) the risk of being affected with the disease or disorder investigated. The association may be a direct effect from a putative causal variant or result from a surrogate polymorphism being in linkage disequilibrium (indirect association) with a disease causing mutation. The type of effect cannot as such be deduced from the association analysis and promising results from association studies are therefore usually only the first step towards a better understanding of an aetiological coherence.

Single-marker analysis is without doubt the most commonly used type of genetic association analysis and many different methods for testing single-markers exists. We will consider the difference between genotype and allele based analysis, and note why the latter generally should not be used. We will briefly describe how a number of genetic models (modes of penetrance) are subspaces of the full genotypic model and includes the genotypic null model of no association. We reflect a bit on how to test single-markers located on autosomes and finally discuss some of the special issues arising when analysing loci on the X chromosome.

The section closes with a small subsection on haplotype analysis and some further notes on the haplotypes in *MBL2* (paper 1 and 2) and how a multiple SNPs model using the six single markers can be re-parameterised as a model using the seven haplotypes.

3.1.1 Single-marker analysis

Genotype-based analysis

Let us consider a locus given by a diallelic marker, e.g. a single nucleotide polymorphism (SNP), and refer to the major (more frequent) and minor (less frequent) alleles by S and s, respectively. Let $\tau = P(s)$ denote the minor allele frequency in the population. We will refer to the three possible genotypes (S/S, S/s and s/s) by G_0, G_1 and G_2 with the indexing indicating number of minor alleles and we let $g_j = P(G_j)$ denote the genotype frequencies in the population. We assume $0 < g_j < 1$ for all $j = 0,1,2$. In the following we will consider a dichotomous trait D and without loss of generality assume that this is a binary disease state: $D = 0$ for subjects without disease (controls) and $D = 1$ for subjects with the disease (cases). Let $0 < \kappa < 1$ denote the prevalence of the disease in the population, $\kappa = P(D = 1)$.

Let π_0, π_1 and π_2 denote the penetrances of the genotypes, i.e. the conditional probabilities of disease given genotype, $P(D = 1|G_j)$. We will assume that $0 < \pi_j < 1$ for all $j = 0,1,2$. Turning this upside down let $\gamma_{ij} = P(G_j|D = i)$, for all $j = 0,1,2$ and $i = 0,1$, denote the conditional genotype frequencies given disease status. As with the other probabilities we will assume $0 < \gamma_{ij} < 1$ for all i and j. The connection between these conditional probabilities is given by Bayes' theorem:

$$P(G_j|D=i) = \frac{P(D=i|G_j)P(G_j)}{P(D=i)}.$$

We see that $\gamma_{0j} = (1-\pi_j)g_j/(1-\kappa)$ and $\gamma_{1j} = \pi_j g_j/\kappa$, for all $j = 0,1,2$. Using the law of total probability we can calculate κ and $(\kappa - 1)$ by

$$P(D=i) = \sum_{j=0}^{2} P(D=i|G_j)P(G_j).$$

If we disregard the potential influence of other factors (including confounding) than that of a single SNP, then the two by three contingency table 3.1 of genotype counts in cases and controls is a useful and reasonable starting point for investigating connection (association) between locus and disease.

Table 3.1 Genotype counts

Phenotype	$G_0(S/S)$	$G_1(S/s)$	$G_2(s/s)$	Total
$D=0$	n_{00}	n_{01}	n_{02}	N_0
$D=1$	n_{10}	n_{11}	n_{12}	N_1
Total	n_0	n_1	n_2	N

Conditional on the number of subjects, N_i, genotype counts within trait (n_{i0}, n_{i1}, n_{i2}) are outcome from multinomial distributions $mult(N_i, (\gamma_{i0}, \gamma_{i1}, \gamma_{i2}))$ for $i = 0,1$ and the two random vectors of genotype counts are independent conditional on $\sum_{j=0}^{2} \gamma_{ij} = 1$ for $i = 0,1$. Thus the maximum likelihood estimate of γ_{ij} is $\hat{\gamma}_{ij} = n_{ij}/N_i$ for all $i = 0,1$ and all $j = 0,1,2$.

The log-likelihood function for a multinomial distribution is

$$l_i((\gamma_{i0}, \gamma_{i1}, \gamma_{i2}); (n_{i0}, n_{i1}, n_{i2})) = \sum_{j=0}^{2} n_{ij} \log(\gamma_{ij})$$

which we can write as a function of two parameters under the restriction that $\sum_{j=0}^{2} \gamma_{ij} = 1$:

$$l_i((\gamma_{i0}, \gamma_{i2}); (n_{i0}, n_{i1}, n_{i2})) = n_{i0}\log(\gamma_{i0}) + n_{i1}\log(1-\gamma_{i0}-\gamma_{i2}) + n_{i2}\log(\gamma_{i2}).$$

Assuming statistical independence between cases and controls the log-likelihood for the four dimensional parameter space of the full (saturated) genotypic model

$$\Gamma_G = \{\gamma | 0 < \gamma_{ij} < 1 \text{ and } \gamma_{i0} + \gamma_{i2} < 1 \text{ for all } i = 0,1 \text{ and all } j = 0,1,2\}$$

will then be

$$l(\gamma; \mathbf{n}) = \sum_{i=0}^{1} l_i((\gamma_{i0}, \gamma_{i2}); (n_{i0}, n_{i1}, n_{i2})),$$

where $\gamma = (\gamma_{00}, \gamma_{02}, \gamma_{10}, \gamma_{12})$ and $\mathbf{n} = (n_{00}, n_{01}, n_{02}, n_{10}, n_{11}, n_{12})$.

Genotypic null model

The genotypic null model is that of equal genotype frequencies between cases and controls, i.e. $\gamma_{ij} = \gamma_j$ for $i = 0, 1$ and all $j = 0, 1, 2$. The null model corresponds to the following subspace of Γ_G

$$\Gamma_{0_G} = \{\gamma | \gamma_{00} = \gamma_{10} \text{ and } \gamma_{02} = \gamma_{12}\} \cap \Gamma_G$$

and the maximum likelihood estimate of γ_{ij} under Γ_{0_G} is $\hat{g}_j = n_j/N$ for all $j = 0, 1, 2$.

From the formulae connecting π_j and γ_{ij} we see that this null model implies $\pi_0 = \pi_1 = \pi_2 = \kappa$, i.e. all three penetrances equals the prevalence of disease in the population. Restrictions on the penetrances π_j are often imposed and correspond to underlying genetic models with parameter spaces between Γ_{0_G} and Γ_G. These models are usually described in terms of restrictions on genotypic relative risks or odds ratios as shown in the next subsection.

Measures of disequilibrium

There are two important measures of disequilibrium of genotype and haplotype frequencies: statistical association between alleles at the same locus (non-independence of chromosomes) and statistical association between alleles on the same chromosome (non-independence of loci). The former of these is known as Hardy-Weinberg (dis)equilibrium (HWE) and the latter is the so-called linkage disequilibrium (LD).

Under the assumption of independence between chromosomes it follows from probability theory that $g_0 = P(S/S) = P(S)^2$, $g_1 = P(S/s) = P(S)P(s) + P(s)P(S) = 2P(S)P(s)$ and $g_2 = P(s/s) = P(s)^2$. When this connection between genotype and allele frequencies is fulfilled we say that we have Hardy-Weinberg proportions. Thus, the measures of Hardy-Weinberg disequilibrium are of the form $D_S = P(S/S) - P(S)^2$. The Hardy-Weinberg principle states that genotype frequencies of a population will assume these proportions after a single generation of random mating. When the random mating assumption is violated, the population will not have Hardy–Weinberg proportions. A common cause of non-random mating is inbreeding, which causes an increase in homozygosity for all genes.

If a population violates one of the following four assumptions, the population may continue to have Hardy–Weinberg proportions in each generation, but the allele frequencies will change over time: 1) *Selection*, in general, causes allele frequencies to change, often quite rapidly. While directional selection eventually leads to the loss of all alleles except the favored one, some forms of selection, such as balancing selection, lead to equilibrium without loss of alleles. 2) *Mutation* will have a very subtle effect on allele frequencies. Mutation rates are of the order 10^{-4} to 10^{-8}, and the change in allele frequency will be, at most, of the same order. Recurrent mutations will maintain alleles in the population, even if there is strong selection against them. 3) *Migration* genetically links two or more populations together. In general, allele frequencies will become more homogeneous among the populations. Some models for migration inherently include non-random mating (Wahlund effect, for example). For those models, the Hardy–Weinberg proportions will normally not be valid. 4) *Small population size* can cause a random change in allele frequencies. This is due to a sampling effect, and is called genetic drift. Sampling effects are most important when the allele is present in a small number of copies.

The corresponding measures of LD are of the form $D_{S_1S_2} = P(S_1S_2) - P(S_1)P(S_2)$, where S_1 and S_2 are alleles at two different loci on the same chromosome. So, LD describes a situation in which a haplotype occurs more (or less) frequently in a population than would be expected by chance, and thus knowledge of an allele at one locus can be used to predict the allele at a second locus. Due to recombinations between loci, LD decay over time.

Allele based analysis

Instead of counting genotypes and thus subjects one could also count chromosomes, i.e. 2 observations per subject. It is obvious how to go from genotype frequencies to allele frequencies whereas the opposite is only possible under assumptions like that of Hardy-Weinberg proportions in the population. Let F denote the inbreeding coefficient in the population:

$$F = 1 - \frac{\text{Observed number of heterozygotes}}{\text{Expected number of heterozygotes under HWE}}.$$

Then there is the following relationship between allele and genotype frequencies:

$$\begin{aligned} g_0 &= (1-\tau)^2(1-F) + (1-\tau)F = (1-\tau)^2 + F(1-\tau)\tau, \\ g_1 &= 2(1-\tau)\tau(1-F), \\ g_2 &= \tau^2(1-F) + \tau F = \tau^2 + F(1-\tau)\tau. \end{aligned}$$

If Hardy-Weinberg equilibrium is present then $F = 0$ and genotype frequencies can be deduced from the minor (or major) allele frequency: $g_0 = (1-\tau)^2$, $g_1 = 2(1-\tau)\tau$ and $g_2 = \tau^2$. The calculation of allele counts from genotype counts is given by the two by two contingency table 3.2.

Table 3.2 Allele counts

Phenotype	S	s	Total
$D = 0$	$a_{00} = 2n_{00} + n_{01}$	$a_{01} = n_{01} + 2n_{02}$	$A_0 = 2N_0$
$D = 1$	$a_{10} = 2n_{10} + n_{11}$	$a_{11} = n_{11} + 2n_{12}$	$A_1 = 2N_1$
Total	$a_0 = 2n_0 + n_1$	$a_1 = n_1 + 2n_2$	$A = 2N$

Allelic null model

The allelic null model assumes equal allele frequencies in cases and controls. If we let $\tau_i, i = 0,1$ denote the conditional allele frequencies of s given disease status then we have the following allelic null model: $\tau_i = \tau$, for all $i = 0,1$. Noting that $\tau_i = \gamma_{i1}/2 + \gamma_{i2}$ and inserting $\gamma_{i1} = 1 - \gamma_{i0} - \gamma_{i2}$ we see that $\tau_i = 1/2 - (\gamma_{i0} - \gamma_{i2})/2$. Thus $\tau_0 = \tau_1$ if and only if $\gamma_{00} - \gamma_{02} = \gamma_{10} - \gamma_{12}$ and the allelic null model is given by the following subset of the four dimensional parameter space Γ_G:

$$\Gamma_{0_A} = \{\gamma | \gamma_{00} - \gamma_{02} = \gamma_{10} - \gamma_{12}\} \cap \Gamma_G.$$

As $\gamma \in \Gamma_{0_G}$ implies $\gamma \in \Gamma_{0_A}$ but not the other way around, we note that $\Gamma_{0_G} \subset \Gamma_{0_A} \subset \Gamma_G$, i.e. that the allelic null model is less stringent than the genotypic null model. We will return to the question and debate of which alternative (allele-based) model to consider against the allelic null model.

3.1.2 Measuring genotypic association with disease

If we use the major allele homozygote genotype as reference then the strength of association between locus and disease may be expressed in terms of the genotype relative risk (GRR):

$$\lambda_j = \frac{\pi_j}{\pi_0}, \quad \text{for all} \quad j = 0,1,2.$$

Note that by construction $\lambda_0 \equiv 1$.Using the connection between π_j and γ_{1j} this can also be written in terms of genotype frequencies by comparing ratios in cases with those of the population:

$$\lambda_j = \frac{\gamma_{1j}/\gamma_{10}}{g_j/g_0}, \quad j = 1,2.$$

Following Joo et al. (2009) and letting $\lambda = (\lambda_1, \lambda_2)$ we can write up the space of the full model determined by the GRRs as

$$\Lambda = \{\lambda | \lambda_1 > 0 \text{ and } \lambda_2 > 0\}.$$

Since $\pi_0 = \pi_1 = \pi_2$ if and only if $\lambda_1 = \lambda_2 = 1$, we see that

$$\Lambda_0 = \{\lambda | \lambda_1 = \lambda_2 = 1\}$$

is equivalent to

$$\Pi_0 = \{\pi | \pi_0 = \pi_1 = \pi_2\}.$$

It is also straightforward to see that Λ is equivalent to

$$\Pi = \{\pi | \pi_0 > 0 \text{ and } \pi_1 > \pi_0 \text{ and } \pi_2 > \pi_0\}.$$

A number of genetic models are easily defined as subsets of Λ, each of which includes Λ_0 and thus defines possible alternative models to this null model. Moreover, these models can also be written as subspaces of the penetrance space Π, each of which includes Π_0:

Recessive	:	$\Lambda_R = \{\lambda \| \lambda_1 = 1 \text{ and } \lambda_2 > 1\}$	or	$\Pi_R = \{\pi \| \pi_1 = \pi_0 \text{ and } \pi_2 > \pi_0\}$
Additive	:	$\Lambda_A = \{\lambda \| \lambda_2 = 2\lambda_1 - 1\}$	or	$\Pi_A = \{\pi \| \pi_2 = 2\pi_1 - \pi_0\}$
Multiplicative	:	$\Lambda_M = \{\lambda \| \lambda_2 = \lambda_1^2\}$	or	$\Pi_M = \{\pi \| \pi_2 = \pi_1^2/\pi_0\}$
Dominant	:	$\Lambda_D = \{\lambda \| \lambda_1 = \lambda_2 > 1\}$	or	$\Pi_D = \{\pi \| \pi_1 = \pi_2 > \pi_0\}$
Overdominant	:	$\Lambda_O = \{\lambda \| \lambda_1 > 1 \text{ and } \lambda_1 > \lambda_2\}$	or	$\Pi_O = \{\pi \| \pi_1 > \pi_0 \text{ and } \pi_1 > \pi_2\}$
Underdominant	:	$\Lambda_U = \{\lambda \| \lambda_1 < 1 \text{ and } \lambda_1 < \lambda_2\}$	or	$\Pi_U = \{\pi \| \pi_1 < \pi_0 \text{ and } \pi_1 < \pi_2\}$

In case one of the alleles (we assume it is the minor) gives rise to a higher risk, then according to Joo et al. (2009), the underlying genetic model is contained in the following subspace:

$$\widetilde{\Lambda} = \{\lambda | 1 \leq \lambda_1 \leq \lambda_2 \text{ and } \lambda_2 > 1\} \cup \{(\lambda_1, \lambda_2) : 0 \leq \lambda_2 \leq \lambda_1 \leq 1 \text{ and } \lambda_2 < 1\},$$

which includes Λ_R, Λ_A, Λ_M and Λ_D, i.e. contains the recessive, additive, multiplicative and dominant models. The full model is given as the following union

$$\Lambda = \widetilde{\Lambda} \cup \Lambda_0 \cup \Lambda_O \cup \Lambda_U.$$

Oftentimes the association is measured by an odds ratio (OR) rather than a GRR, i.e. as a ratio between odds of disease $\pi_j/(1-\pi_j)$:

$$\theta_1 = \frac{\pi_1}{1-\pi_1} / \frac{\pi_0}{1-\pi_0} \text{ and } \theta_2 = \frac{\pi_2}{1-\pi_2} / \frac{\pi_0}{1-\pi_0}.$$

3.1.3 Testing autosomal single-markers (SNPs)

Using likelihood inference the genotypic model Γ_G can be tested as the alternative to the genotypic null model Γ_{0_G} by the following likelihood ratio test:

$$-2\log(Q) = 2\sum_{i=0}^{1}\sum_{j=0}^{2} n_{ij} \log(\frac{n_{ij}}{N_i n_j/N}).$$

This test statistic is found by inserting likelihood ratio estimates $\hat{\gamma}_{ij} = n_j/N$ under $H_{0_G} : \gamma \in \Gamma_{0_G}$ into the multinomial likelihood function and finding the ratio Q between this and the corresponding likelihood under $H_G : \gamma \in \Gamma_G$, i.e. with $\hat{\gamma}_{ij} = n_{ij}/N_i$ inserted. Finally, minus two times the natural logarithm of Q is calculated to obtain a test statistic for which the asymptotic distribution have been deduced and equals a χ^2-distribution on 2 degrees-of-freedom, χ^2_2. Under the same conditions needed for this asymptotic result, this test statistic is also approximately equal to Pearson's χ^2-test for independence between row and column variables in the two-by-three contingency table 3.1:

$$X^2_{Pearson} = \sum_{i=0}^{1}\sum_{j=0}^{2} \frac{(n_{ij} - N_i n_j/N)^2}{N_i n_j/N}.$$

However, the null model Γ_{0_G} is also often tested against other alternatives implied by the genetic models mentioned above. If such a genetic model is reasonably correct, the model-based tests may be more powerful than the genotypic test which on the other hand is more robust to model misspecification as it includes the other models. The study by Joo et al. (2009) (see also the commentary by Zheng et al., 2009) concludes that there is no single best test for case-control association studies and that robust tests which do not depend on specific assumptions of genetic models are preferable. Their results indicate that Pearson's χ^2_2-test performs equally well as the tests that combine trend tests corresponding to specific recessive, additive and dominant genetic models. Balding (2006) also noted that χ^2_2 or Fisher's exact test have reasonable power regardless of the underlying model but are less powerful when the risks are additive, than tests tailored to this scenario. Therefore, unless the genetic model is known or may be restricted to be between recessive and dominant model (precluding e.g. the over- and underdominant models), either χ^2_2 or the MIN2 (Wellcome Trust Case Control Consortium, 2007; Joo et al., 2009) should be considered. The MIN2 test is obtained as the minimum of the two p-values from an additive trend test and from the χ^2_2-test. This minimum is not in itself a valid p-value, but Joo et al. (2009) derived the asymptotic p-value and some threshold values for a few significance levels.

An often used class of tests is called Cochran-Armitage Trend Test (CATT) and dates back to Cochran (1954) and Armitage (1955). To define a specific CATT, increasing scores $x_0 \leq x_1 \leq x_2$ and $x_0 < x_2$ or decreasing scores $x_0 \geq x_1 \geq x_2$ and $x_0 > x_2$ are assigned to the genotypes (G_0, G_1, G_2), e.g. corresponding to the following models: recessive (0,0,1), additive (0,1,2) and dominant (0,1,1). As the CATT is invariant to a linear transformation of the scores, $\mathbf{x} = (x_0, x_1, x_2)$ can be replaced by $\mathbf{x} = (0, \eta, 1)$ with $\eta = (x_1 - x_0)/(x_2 - x_0)$ (c.f. Zheng et al., 2003). So for the three examples, the recessive and dominant scores remain unchanged while for the additive model we obtain the scores (0,1/2,1). Zheng et al. (2003) showed that these scores are optimal for the recessive and dominant models with respect to minimised sample size to achieve prespecified type I error and power (or type II error), but only locally optimal for the additive model. Furthermore they showed that the multiplicative model is asymptotically equivalent to the additive model and that the scores used for the additive model (0,1/2,1) are locally optimal for the multiplicative model, i.e. the same trend test (set of scores) can be used for both models.

Given the score vector $\mathbf{x}$ and using formulae corresponding to those in Joo et al. (2009), the CATT statistic is given by

$$Z_{CATT}(\mathbf{x}) = \frac{U(\mathbf{x})}{\sqrt{\widehat{Var}_{\Gamma_{0_G}}(U(\mathbf{x}))}}$$

or equivalently by $X^2_{CATT}(\mathbf{x}) = (Z_{CATT}(\mathbf{x}))^2$, where

$$U(\mathbf{x}) = \sqrt{N}\sum_{j=0}^{2} x_j(\hat{\gamma}_{1j} - \hat{\gamma}_{0j})$$

and

$$\widehat{Var}_{\Gamma_{0_G}}(U(\mathbf{x})) = \frac{N^2}{N_0N_1}\{\sum_{j=0}^{2} x_j^2\hat{g}_j - (\sum_{j=0}^{2} x_j\hat{g}_j)^2\}.$$

Here $\hat{\gamma}_{ij} = n_{ij}/N_i$ are maximum likelihood estimates for γ_{ij} under $\gamma \in \Gamma_G$ and $\hat{g}_j = n_j/N$ are maximum likelihood estimates for γ_{ij} under $\gamma \in \Gamma_{0_G}$. Under Γ_{0_G} and for a fixed score vector $\mathbf{x}$, the asymptotic distribution of $Z_{CATT}(\mathbf{x})$ is a standard Gaussian, $N(0,1)$, and consequently the asymptotic distribution of $X^2_{CATT}(\mathbf{x})$ is a chi-squared distribution with one degree-of-freedom, χ^2_1.

In Zheng et al. (2009) it was shown that if we allow the scores to be unordered then $X^2_{Pearson}$ is also a trend test with score vector $\mathbf{x} = (n_{10}/n_0, n_{11}/n_1, n_{12}/n_2)$ or equivalently $\mathbf{x} = (0, \eta, 1)$ where

$$\eta = \frac{n_{11}/n_1 - n_{10}/n_0}{n_{12}/n_2 - n_{10}/n_0} = \frac{(n_{11}n_0 - n_{10}n_1)n_2}{(n_{12}n_0 - n_{10}n_2)n_1}.$$

If the true scores are ordered, then for some fixed $\eta \in [0,1]$ (or the equivalent $\mathbf{x}$) the corresponding (optimal) CATT with one degree-of-freedom will necessarily be more powerful than Pearson's χ^2-test for which we use two degrees-of-freedom. On the other hand, allowing the scores to be unordered imply the robustness of Pearson's χ^2 when the true model of inheritance is not between the recessive and dominant.

Very recently, Loley et al. (2013) proposed a unifying framework using the generalised linear model (GLM) to encompass both robust testing for association by use of a MAX test (maximum of dominant, additive and recessive tests) and selection of the best fitting genetic model. Their approach can be used both for usual case-control data, for family-based data and for matched pairs data, e.g. matched case-control data. By utilising the GLM framework it is also possible to consider continuous traits, count data, categorical traits and survival data. Furthermore, the inclusion of covariates is possible.

The allele-based 2x2 contingency table test: should it be used or not?

In allele-based analyses the null hypothesis of interest is that of no difference in allele frequencies between cases and controls, $H_0^A : \tau_0 = \tau_1 = \tau$. Under H_0^A the maximum likelihood estimate is $\hat{\tau} = (n_1 + 2n_2)/(2N) = a_1/A$ (see table 3.2). Note that $1 - \hat{\tau} = a_0/A$. This may then be tested using Pearson's χ^2-test given by

$$X_A^2 = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{(a_{ij} - A_ia_j/A)^2}{A_ia_j/A}$$

evaluated by a χ^2-distribution on one degree-of-freedom. However, as noted by Sasieni (1997) this distributional assumption for X_A^2 relies on binomial distributions of the test statistics which is only reasonable under an assumption of statistically independent alleles, i.e. under an assumption of HWE of genotypes in the combined population.

Sasieni (1997) compared systematically the allele- and genotype-based chi-square tests and concluded that these are (asymptotically) equivalent if and only if HWE holds in the combined sample and that X_A^2 is otherwise invalid and thus should not be used. This asymptotic equivalence holds both under the null hypothesis of no association and under the alternative hypotheses, c.f. Zheng (2008). Schaid et al. (1999) quantified the bias of the expected type-I error rate under the null hypothesis of no association when there are deviations from HWE and also devised a correction to account for the deviations. The recommendation not to use the allele-based test (Sasieni, 1997) has not always been followed, as noted by Guedj et al. (2008), Zheng et al. (2009) and Izbicki et al. (2012).

The asymptotic equivalence under the null hypothesis was complemented and more firmly proved by Guedj et al. (2008) when HWE holds in the population, and was noted by Knapp (2008) to also follow easily from results in Knapp (2001). Zheng (2008) showed that the equivalence under the null hypothesis is not true under the alternative hypotheses except under certain conditions in addition to HWE in the population, e.g. if also the proportion of cases in the case-control sample is an unbiased estimate for the disease prevalence. Zheng (2008) concludes that for the additive model, the trend test is always more powerful than the allelic test but that the opposite is the case in some scenarios under the recessive, dominant and multiplicative models. As an alternative to the asymptotic-based tests, Guedj et al. (2006) proposed an unbiased exact allelic test which has been implemented in the R package allelelic.

Recently Wang (2012) proposed a new framework for testing equality of allele frequencies between cases and controls while allowing for deviations from HWE. Furthermore, this method does not depend on a specific genetic disease model and the test statistics are evaluated on one degree-of-freedom. The method of Wang (2012) was implemented in the R package iGasso. In another recent study, Izbicki et al. (2012) found that the genotype-based hypothesis was not rejected while the allele-based was rejected even after devising an allelic test, similar to those in Wang (2012), which should be appropriate also under deviations from HWE. To circumvent this incoherence they suggested a Bayesian approach. It is not totally clear to us, though, if this incoherence really stems from the difference in dimension (as suggested by the authors) or might just result from a difference in power.

In conclusion, there is no good reason to use the original allele-based Pearson's chi-square test. There may be reasons to used either the exact test by Guedj et al. (2006) or the framework recently proposed by Wang (2012).

3.1.4 Analysis of markers on the sex chromosomes

Since the number of X and Y chromosomes carried by males and females normally differ (males: 1 X and 1 Y; females: 2 X and 0 Y) analyses of genetic markers on the sex chromosomes may require different methods than those usually applied in autosomes (Clayton, 2008; Clayton, 2009) where both gender (normally) carries two homologous chromosomes.

The Y chromosome, which is now approximately 58 Mb, has been progressively degraded during evolution and though originating from an ancestral homologous chromosome pair (Ohno, 1967), now consist largely of repeated sequences. In contrast, almost all original genes are conserved on the X chromosome (Mangs et al., 2007) which is about 155 Mb long. The pseudoautosomal regions (PAR) PAR1 and PAR2 at the telomeres of respectively the short (p) and long (q) arms of the Y chromosome, recombines during meiosis with the corresponding regions of the X chromosome and behaves like autosomes whereas the other part, the male-specific regions (MSY), behaves differently (see e.g. Skaletsky et al., 2003). The PAR1 is 2.6 Mb and appears important for male fertility in contrast to the much shorter PAR2 of only 320 kb, see Mangs et al. (2007) for a review of the PAR. The MSY comprises 95% of the length of the Y chromosome and is a mosaic of heterochromatic and euchromatic sequences with the latter being divided into three classes: X-transposed, X-degenerate and ampliconic (Skaletsky et al., 2003). Here it should be noted that the X-transposed sequences in Yp11.2 (referred to as the X-transposed-region, XTR, by Veerappa et al., 2013), comprising 2 blocks of totally 3.4 Mb, are 99% identical to the DNA sequences in chromosome Xq21.3 and is a result of duplication and X-to-Y transpositions occurring after the divergence of the human and chimpanzee lineages (Skaletsky et al., 2003; Veerappa et al., 2013). However, according to Skaletsky et al. (2003) there is no X-Y crossover during male meiosis in the XTR.

Biologically, some compensatory mechanism is needed to retain balance between males and females for genes on X that are lost on the Y. Inactivation of one of the female X chromosomes early during development seems to be an accepted mechanism resulting in inactivation of one of the alleles per female locus (Augui et al., 2011). In the main, inactivation is believed to happen randomly such that approximately half of the female cells have a paternally derived inactive X chromosome while the other half is maternally derived (Amos-Landgraf et al., 2006). The genes in PAR1 escape X inactivation whereas at least one gene in PAR2 is subject to both X and Y inactivation (Mangs et al., 2007). In the XTR some of the genes also escape inactivation, c.f. Veerappa et al. (2013) who recently suggested the existence of another PAR, named PAR3, located in the XTR 700 kb from the boundary of PAR1. Generally, Ohno's hypothesis of dosage compensation from 1967 (Ohno, 1967; Xiong et al., 2010) have been accepted as the result of an evolutionary process where X-linked genes show approximately double the expression of autosomal genes per active allele to compensate for the inactivation of roughly half of the alleles. Nevertheless, Xiong et al. (2010) suggests that this may be an artefact of the methods used and that results from RNA sequencing data rejects Ohno's explanation of dosage compensation and requests new theories to explain observed between-gender dosage compensation. The hypothesis of inactivation to achieve dosage compensation was not challenged though—only the evolutionary model behind.

Analysis of *MAOA*

MAOA, located on Xp11.3, which was investigated in the suicide study (paper 4) is clearly not close to the usual pseudoautosomal (PAR1 and PAR2) or the XTR (and PAR3) parts of the X chromosome. Therefore analysis of the *MAOA*uVNTR marker may require a different approach than those usually applied to autosomal loci (Clayton, 2008; Clayton, 2009). Although *MAOA* is far from the PAR and XTR one may still pose the question if *MAOA* belongs to the set of exceptions? Though earlier reports have suggested the contrary, a study using rodent/human somatic cell hybrids suggested that *MAOA* escape X inactivation (Carrel et al., 2005). Also in a study on mRNA expression in postmortem brain tissues, *MAOA* methylation ratios for the 3- and 4-repeat alleles of *MAOA*uVNTR (referred to as pVNTR in that study) did not correlate with inactivation ratios and thus called upon an alternative explanation of *MAOA* dosage compensation in females (Pinsonneault et al., 2006). Nevertheless, the study by Stabellini et al. (2009) concluded that *MAOA* is subjected to X inactivation in normal human fibroblasts.

Under the assumption of inactivation, the effect of the minor allele in males has to be equivalent to the difference between the two homozygote genotypes in females (Clayton, 2009) or in other words; male carriers of the minor allele should correspond to female homozygote carriers (Clayton, 2008). This was done by coding an additive term A_i to be 0 or 2 for male X chromosomal loci, whereas a corresponding dominance term, D_i, was always set to 0. In this setting, female genotypes were coded as for autosomal chromosomes. We could equivalently choose to divide female allele counts by two, i.e. $A_i \in \{0, 0.5, 1\}$ for females and $A_i \in \{0, 1\}$ for males. However, we chose the former parametrisation in paper 4 (Buttenschøn et al., 2013).

Obviously, only females can contribute to the dominance part, but if one allele is active what is the interpretation of the dominance effect then? Of course, for homozygous females we know the allele with certainty whereas for heterozygous females we will have only approximately a 50/50 chance to choose the active allele. If we assume instead that the region (e.g. *MAOA*) analysed escape inactivation then male subjects should be coded 0 or 1 for the additive term A_i. In paper 4, a combined 2 degrees-of-freedom chi-squared test was calculated under both inactivation scenarios by adding the two 1 degree-of-freedom chi-squared test statistics for the additive and dominance effects from conditional logistic regressions stratified on gender and age-group.

3.1.5 Haplotype analysis

Haplotype analysis and other multilocus approaches may increase the power to detect disease association but generally introduce also the problem of determining the gametic phase. If haplotypes are identifiable, i.e. the phase is (assumed) known or can be guessed with almost certainty, then the same methods as used for single-markers can be used by treating each haplotype as if it was a SNP. Usually, the linkage phase of haplotypes cannot be unambiguously determined and other methods taking this uncertainty into account should be used. An example of such a method is the score-method by Schaid et al. (2002). For an application we refer to paper 5 (Buttenschøn et al., 2010) where we used a sliding window approach of this score method for haplotype analyses of two, three and four consecutive markers. The method by Schaid et al. (2002) has been implemented in the R package haplo.stats.

In papers 1 and 2 we analysed multilocus genotypes and haplotypes, but phase was assumed known as linkage disequilibrium implies only few haplotypes to be commonly observed using these markers, see below. However, the validity of the identified haplotypes was also checked by inferring phased haplotypes from genotypes with the software BEAGLE (Browning et al., 2007). See section 3.3 for more on BEAGLE and and other imputation methods.

MBL2 haplotypes

As noted in subsection 2.2.1, only seven haplotypes are usually observed from the six SNPs in *MBL2*: HYPA, LYPA, LYQA, LXPA, HYPD, LYPB and LYQC. Let us in this subsection refer to the six SNPs H/L (rs11003125), X/Y (rs7096206), P/Q (rs7095891), A/D (rs5030737), A/B (rs1800450) and A/C (rs1800451) by by their minor allele: H, X, Q, D, B and C, respectively. The latter three are positioned in *MBL2* exon 1 (see figure 4.1 or Figure 1 in paper 1) and have never been observed in the same chromosome, i.e. these can be seen as one 4-allelic marker and explains why the haplotypes can be represented by just 4 alleles. Formally these would be HYPAAA, LYPAAA, LYQAAA, LXPAAA, HYPDAA, LYPABA and LYQAAC. We code genotypes and multilocus genotypes by the number of minor alleles, i.e. 0 (M/M), 1 (M/m) and 2 (m/m) with M and m denoting major and minor allele, respectively. Obviously, each individual carry exactly two of the haplotypes: 0, 1 or 2 of each. Thus, with the 0/1/2, coding the sum of multilocus genotype codes over the seven possible haplotypes will be exactly two (or missing) for each individual. Denoting the counts by the corresponding allele or haplotype, the following set of seven equations thus must be true:

$$\begin{aligned}
H &= HYPA+HYPD = 2-(LYPA+LYQA+LXPA+LYPB+LYQC) = 2-L\\
X &= LXPA\\
Q &= LYQA+LYQC\\
D &= HYPD\\
B &= LYPB\\
C &= LYQC\\
2 &= HYPA+LYPA+LYQA+LXPA+HYPD+LYPB+LYQC.
\end{aligned}$$

A direct calculation now gives that the model for multiple SNPs with a trend parameter for each of the six single markers

$$\text{logit}(p) = \alpha_0+\alpha_1 H+\alpha_2 X+\alpha_3 Q+\alpha_4 D+\alpha_5 B+\alpha_6 C,$$

can be re-parameterised to the model containing a trend parameter for each of the seven haplotypes

$$\text{logit}(p) = \beta_0+\beta_1 LYPA+\beta_2 LYQA+\beta_3 LXPA+\beta_4 HYPD+\beta_5 LYPB+\beta_6 LYQC,$$

with

$$\begin{aligned}
\beta_0 &= \alpha_0+2\alpha_1\\
\beta_1 &= -\alpha_1\\
\beta_2 &= \alpha_3-\alpha_1\\
\beta_3 &= \alpha_2-\alpha_1\\
\beta_4 &= \alpha_4\\
\beta_5 &= \alpha_5-\alpha_1\\
\beta_6 &= \alpha_6+\alpha_3-\alpha_1.
\end{aligned}$$

3.2 Simulation of G×E genotypic data

In this section we provide further details on the simulation of samples for the G×E simulation study (paper 3). To ensure realistic correlation between the SNPs, we based the simulations on an initial population consisting of 993 unrelated subjects obtained by merging all 11 HapMap3 populations (International HapMap 3 Consortium, 2010) and used phased genomic data from the HapMap3 database[14]. Using a Wright-Fisher forward-time simulation with mutation and recombination, the initial population was then expanded linearly for 500 non-overlapping generations to obtain a base population of 50,000 individuals. Approximated by the harmonic mean of census sizes in each generation (Wright, 1938; Crow et al., 1970) the expected effective population size of the base population was

$$N_e = \frac{N_{gen}}{\sum_{i=1}^{N_{gen}} \frac{1}{N_i}} = 12{,}658,$$

with N_{gen}=500 and N_i the population size for population $i = 1,\ldots,500$. To find N_i simply use $N_i = N_0+\beta i$ with $\beta = \frac{50{,}000-N_0}{N_{gen}}$ where N_0=993 is the size of the initial population. We note that N_e=12,658 seems reasonably for the present Caucasian populations. Further details on the simulation procedure is given in the subsections below.

[14] ftp://ftp.ncbi.nlm.nih.gov/hapmap/phasing/2009-02_phaseIII/HapMap3_r2/

3.2.1 simuGEMS

We used simuPOP (Peng et al., 2005) for the simulations by writing our own Python scripts. We will call this modified collection of simuPOP-based Python scripts the simuPOP-based Gene-Environment Model Simulator (simuGEMS). It is intended for use in a high-performance computing (HPC) environment and graphical user interface (GUI) functionalities have therefore been removed from the scripts. In the present study, we used the GenomeDK hub established by the Genome Denmark project[15].

We have been borrowing massively from the scripts described in Peng et al. (2010) and Peng et al. (2012), which are available as simuGWAS under the menus ”Complete Scripts” and ”Wiley Book” in the simuPOP online cookbook[8]. These scripts were then revised to our needs; specifically we simulated with neutral selection on all SNPs and instead of using trajectory sampling to ensure specific MAFs of predefined DPLs (*see* sbsection 2.2.2), we picked at random among SNPs having a MAF of a certain size in what we refer to as the *base population* (see below). We used default values from simuGWAS with respect to mutation rate (1e-8) and recombination intensity (1e-8). The actual recombination rate is this intensity multiplied by the physical distance in base pairs between adjacent loci. All non-DPLs chosen for the base sample serves as background noise but may obviously be associated with disease as a result of linkage disequilibrium. In principle, any number of non-predisposing genetic and environmental factors can be simulated and included.

The affection status was generated by use of modified scripts from the simuPOP-based Gene-Environment iNteraction Simulator v.2, GENS2[16] (Pinelli et al., 2012). This is used to control the penetrance while allowing for G×E interaction between up to two DPLs and one disease predisposing environmental variable (DPE)—and with the possibility to also include epistasis (G×G interaction). For a number of reasons, we had to correct and revise GENS2 to make it run (see subsection 3.2.2) and to meet some of the scenarios we have in mind. We extended with possibilities to choose binomial (including binary) and multinomial DPE distributions, rather than the Gaussian implemented in GENS2. In principle, the methods described by (Pinelli et al., 2012) may be used for higher order interactions as well, but we believe that e.g. optimisers have to be chosen differently if the dimensionality increases markedly. We have not investigated this practical aspect further though.

To generate case-control samples, we implemented the rejection sampling method used in simuGWAS (Peng et al., 2010). This sampling was carried out by repeatedly generating offsprings (one at a time) from the base population by random mating and determining the offsprings affection status by comparison of the subjects’ disease risk (penetrance) with a random number from the uniform distribution on the interval $[0,1]$, see subsection 3.2.3. The offspring was then classified as affected/unaffected if the penetrance was higher/lower than this uniform random number. Offsprings with a given affection status were rejected if the desired number of subjects with that affection status was already reached. This rejection sampling algorithm is needed as the proportion of affected from the base population will usually be too low for random sampling from the base population to be feasible. Therefore, the original sampling procedure in GENS2 seems inappropriate at least for diseases of low prevalence. Alternatively, a much larger base population would be needed, but we were unable to do this without increasing the effective population sizes as well—and thus presumably inducing LD patterns incompatible with present day human populations. To generate a larger base population without sacrificing the correct

[15] http://genome.au.dk/

[16] http://sourceforge.net/projects/gensim/

effective population size, we would need a larger initial (founder) population. Data from the 1000 Genomes Project[17] (The 1000 Genomes Consortium et al., 2010) might be a source in this respect, as could the increasing number of high coverage sequencing data being generated currently in many studies. In principle any number of non-predisposing genetic and environmental factors can be simulated and included but we decided to omit this extra source of noise at this time. A small set of scenarios were chosen by varying MAF and effect size of DPLs, and prevalence and effect sizes of the DPE. We chose to fix the sample size at 10,000 individuals with 5,000 affected subjects (cases) and 5,000 unaffected (controls).

3.2.2 Changes made to make GENS2 run

We will here briefly describe the most important changes made to get GENS2[16] up and running. The first modification was due to the fact that *integrate.Inf* no longer exist in scipy. Instead we used *numpy.inf* from numpy, which is now a core package of the *SciPy Stack*[18], and corrected accordingly in the Python script file *GensDistribs.py*. Another adjustment needed for the scripts to run were a removal of a comment tag (#) in line 499 (solver='ralg') in *gens.py* and commenting out instead line 497 and 498. The *ralg* algorithm is from OpenOpt[19] and is used during the optimisation of epistasis parameters. Furthermore, we had to comment out lines 515, 517 and 518 of *gens.py* as these caused an "invalid index to scalar variable" error.

We should note that we modified version 2.0-beta and that a new apparently stable (non-beta) version 2.3 was made available 22 November 2013. We have not tested if the problems were fixed but they note that the distribution was tested with the numpy version 1.6.0, scipy version 0.9 and openopt version 0.38. These are rather old versions: numpy version 1.6.0 is from May 2011 (the newest version is 1.8.0 from October 2013); scipy version 0.9 is from March 2011 (current version is 0.13.3 from February 2014); openopt version 0.38 is from March 2012 (current version is 0.52 from December 2013). We therefore anticipate, that the problems mentioned above might still remain in this new version of GENS2.

3.2.3 Penetrance modelling

A mathematical approach named the Multi-Logistic Model (MLM) was suggested by the authors of GENS (Amato et al., 2010) for calculations of penetrances. The MLM method is applying a system of logistic functions to describe disease risk in simulated case-control samples, with different logistic functions for each combination of genotypes. The method is general and allows in principle any number of DPLs and DPEs, and any order of interaction between these variables.

Using the so-called Knowledge Aided Parameterization System (KAPS) (Amato et al., 2010) for one DPL and one DPE or KAPS version 2 (KAPS2) (Pinelli et al., 2012) for two DPLs and one DPE, user input values of certain biological and epidemiological parameters are translated to the coefficients of the MLM, which corresponds in essence to penetrances of the various combinations of DPL (multilocus) genotypes as a function of the DPE. The elements consist of the expected disease prevalence m (the proportion we want in the study sample), the name (id) of DPLs (one or two), allele frequencies of the DPLs in the base population (calculated automatically), relative risk (RR) of the high risk homozygote compared with the low risk homozygote (expected risk ratio), a dominance parameter $W \in [0, 1]$, and the parameters of

[17] http://www.1000genomes.org/

[18] http://www.scipy.org

[19] http://openopt.org

the environmental variable plus the effect in terms of odds ratio of a one-unit increase in the exposure for the (two-locus) genotype conferring the highest risk. The dominance parameter determines the relative risk of the heterozygote genotype as RR^W so that $W = 0$ corresponds to a dominant model, $W = 1$ is a recessive model, and otherwise a co-dominant model is obtained. Over-dominance ($W > 1$) cannot be modelled at present. Furthermore, if two DPLs are provided, the models for G×G and G×E needs to be specified and chosen. There are four possible G×E models to choose from: GEN (no environmental effect), ENV (no genetic effect), GEM (G×E interaction), ADD (additive model, i.e. no interactions), and we added a fifth: NON (no effects). Moreover, it is possible to specify a G×G (epistasis) model in terms of percentages variations on the risk for a maximum of three (out of the nine possible) two-locus genotypes.

It should maybe be noted that the genotype frequencies of each DPL is calculated from the allele frequencies under the assumption of Hardy Weinberg proportions—this is not noted in either of the papers describing GENS and GENS2. However, as the simulations use random mating in non-overlapping generations, this assumption is reasonable.

Let us assume the two DPLs A and B are SNPs with the following genotypes:

$$g_a \in GA = \{AA, Aa, aa\} \text{ and } g_b \in GB = \{BB, Bb, bb\}.$$

We will assume that a and b are the minor alleles and that these are disease predisposing. Furthermore, conditional on affection status the DPLs are assumed independent, i.e. not in linkage disequilibrium. The value of the DPE will be denoted x and its domain will be denoted X. Now the conditional probability of being affected, i.e. the disease risk, given the two-locus genotype (g_a, g_b) and level of exposure x is assumed to be of the following logistic form:

$$P(\text{affected}|g_a, g_b, x) = \frac{1}{1+\exp(\alpha_{(g_a,g_b)} + \beta_{(g_a,g_b)}x)},$$

where $\alpha_{(g_a,g_b)}$ and $\beta_{(g_a,g_b)}$ are the parameters that need to be determined for the desired features to be modeled. Integrating out the DPE, we obtain the total risk given two-locus genotype:

$$TR_{(g_a,g_b)} = P(\text{affected}|g_a, g_b) = \int_X \frac{f_E(x)}{1+\exp(\alpha_{(g_a,g_b)} + \beta_{(g_a,g_b)}x)}dx,$$

where f_E denotes the density of the DPE (this is denoted P^E in Pinelli et al., 2012). In case of a categorical or discretised DPE with values $X_E = \{x_1, \dots, x_k\}$, this formula would instead be:

$$P(\text{affected}|g_a, g_b) = \sum_{x_e \in X_E} \frac{P_E(x_e)}{1+\exp(\alpha_{(g_a,g_b)} + \beta_{(g_a,g_b)}x_e)},$$

where P_E denotes exposure probability and we assume that $\sum_{x_e} P_E(x_e) = 1$ is fulfilled.

The integral (or sum) equation above is defining the system of 9 equations that have to be solved when there are two DPLs, one for each pair of values $(\alpha_{(g_a,g_b)}, \beta_{(g_a,g_b)})$. It is solved numerically and since the number of equations to solve grows exponentially in the degree of the included gene-gene interactions, this may become computationally intractable or unsolvable if higher order interactions between additionally DPLs are considered. Furthermore, restrictions are needed to avoid infinite solutions (Pinelli et al., 2012) and we will now state those implemented in GENS2 for each of the possible models.

First, one of the $\beta_{(g_a,g_b)}$ coefficients is fixed and denoted β_{AB}. This coefficient which corresponds to the logarithm of the odds ratio related to a one unit increase of the DPE is to be chosen by the user (in the option '--OR') as the parameter for the two-locus genotype with highest risk. Next, all $\alpha_{(g_a,g_b)}$ are assumed to be non-zero, i.e. $\alpha_{(g_a,g_b)} \neq 0$ for all $(g_a,g_b) \in GA \times GB$. Now there are five possible models to choose from in simuGEMS:

GEN Genetic effect model (no effect of DPE): $\beta_{(g_a,g_b)} = 0$ and not all $\alpha_{(g_a,g_b)}$ equal, i.e. there exists $(g_a,g_b),(g_c,g_d) \in GA \times GB$ such that $\alpha_{(g_a,g_b)} \neq \alpha_{(g_c,g_d)}$.

ENV Environmental effect model (no effect of DPLs): $\alpha_{(g_a,g_b)} = \alpha_{(g_c,g_d)}$ and $\beta_{(g_a,g_b)} = \beta_{AB}$ for all $(g_a,g_b),(g_c,g_d) \in GA \times GB$.

GEM Genetic modulation of environmental effect: $\alpha_{(g_a,g_b)} = \alpha_{(g_c,g_d)}$ and $\beta_{(g_a,g_b)} \neq 0$ for all $(g_a,g_b),(g_c,g_d) \in GA \times GB$.

ADD Additive polygenic and environmental model: $\beta_{(g_a,g_b)} = \beta_{AB}$ for all $(g_a,g_b) \in GA \times GB$.

NON Null model (which we added—not included in GENS2): $\alpha_{(g_a,g_b)} = \alpha_{(g_c,g_d)}$ and $\beta_{(g_a,g_b)} = 0$ for all $(g_a,g_b),(g_c,g_d) \in GA \times GB$.

Note that the GEM model impose a modelling without main effects of the DPLs. If all $\beta_{(g_a,g_b)}$ equals β_{AB} in GEM this would be the same model as the ENV so an additional reasonable constraint not stated in Pinelli et al., 2012 is that they are not all equal, i.e. that there exists $(g_a,g_b),(g_c,g_d) \in GA \times GB$ such that $\beta_{(g_a,g_b)} \neq \beta_{(g_c,g_d)}$. Finally, the epistasis (G×G) model needs to be defined in terms of a 3×3 matrix of changes to the penetrances for each two-locus combination of the DPL genotypes (for further description see Pinelli et al., 2012).

3.3 Imputation of SNPs

A huge problem for psychiatric genetics and genetics of other complex diseases until recently have been underpowered sample sizes. Expectations to find loci of major aetiological impact and corresponding large association effects were never really met. Therefore forces have been joined by pooling data from various projects to increase the power. However, genotyping technology is evolving quickly and genetic markers (typically SNPs) investigated in one project may differ, sometimes largely, from the markers used in other projects. Meta-analysis techniques relying on test statistics or p-values have also been a way to combine the efforts.

In relation to genetics, the term imputation describes the process of assigning genotypes for markers that are either not genotyped or missing by other means in the study sample. There are various reasons why we may want to impute:

1. Obtaining a joint set of genotyped (or imputed) markers across studies and genotyping platforms to enable combined analyses.
2. Imputing unobserved markers to facilitate comparison with earlier findings and to justify replication conclusions.
3. Fine-mapping and maybe resolving potentially causal nonsynonymous variants.
4. Expanding the set of markers to increase the power of the study.
5. Last but not least we may use imputation to fill out missing genotypes.

To do this, a reference panel of haplotypes at a dense set of SNPs is used, for example by using data from the HapMap Project (International HapMap Consortium, 2005; International HapMap

Consortium et al., 2007; International HapMap 3 Consortium, 2010) or the 1000 Genomes Project (The 1000 Genomes Consortium et al., 2010). The imputation methods attempt to identify sharing between reference haplotypes and the underlying haplotypes of the study subjects. This sharing is then used to impute the missing alleles. Essentially there is a phasing step where the haplotypes of each individual are modeled as a mosaic of reference haplotypes. Thereafter, missing genotypes are imputed using the matching haplotypes from the reference set. Of course, with uncertainty resulting in a probability distribution over all three possible genotypes—and maybe even allowing that these probabilities do not add up to one, i.e. allowing for a ”missing genotype” category.

We refer to Marchini et al. (2010) for a review of some of the most prevailing imputation methods in use, including IMPUTE v. 2 (IMPUTE2)[20] (Howie et al., 2009), MACH[21] (Li et al., 2010b) and BEAGLE[22] (Browning et al., 2007).

In the subsections below we will present some unpublished observations and experiences we obtained from using IMPUTE2 with the intention to obtain complete data for logic regression interaction analysis. We decided to use IMPUTE2 for the following reasons:

1. It is fairly fast.
2. It seems to be the most accurate—though differences are small (Marchini et al., 2010).
3. It can impute in specified intervals.
4. It can take two reference panels in the same run.
5. It is possible to let the program choose a ”custom” reference panel for each individual.

The last point means that the program chooses a subset of haplotypes from a larger collection of reference haplotypes thereby speeding up the imputation while maintaining the better accuracy from the larger panel (Howie et al., 2011). This algorithm was implemented in the original IMPUTE2 (Howie et al., 2009) and removes some of the responsibility to choose reference population (which may not always be obvious) while at the same time choosing from a worldwide reference panel (see below). Moreover, IMPUTE2 was found to attain higher accuracy than BEAGLE and even doing so with a shorter computation time (Howie et al., 2011).

As already noted, we have earlier used BEAGLE to check the validity of the haplotypes in paper 1 and 2. We furthermore have used BEAGLE with some success to impute missing genotypes (Foldager et al., 2010). Using a posterior probability threshold of 90%, we were able to recover all missing genotypes for about half of the subjects that had a least one unknown genotype. Genotypes with maximum posterior probability below 90% remained missing and subjects carrying these still had to be excluded when using methods demanding complete data, e.g. logic regression and other machine learning methods. Instead of excluding individuals who miss at least one genotype it would be better if the full information could be used. Decreasing the posterior probability threshold would of course also increase the sample size but this would be at the expense of lowering the genotyping quality and therefore the reliability of the results. Some downstream analysis methods use genotype posterior probabilities as input in place of the genotypes, and this is probably the most efficient way to use the information obtained from imputation.

[20] http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

[21] http://www.sph.umich.edu/csg/abecasis/MACH/

[22] http://faculty.washington.edu/browning/beagle/beagle.html

3.3.1 The problem of missing genotypes

If we are only testing single-markers it is not a big problem to miss out a few individuals. However, when using e.g. machine learning methods complete data is a demand. That is, we would have to discard individuals for which genotypes are missing for at least one SNP among the markers investigated. To show how big a problem this may be, let us consider data from the Danish Genomic Medicine for Schizophrenia (GEMS) project (Hollegaard et al., 2011). Genotyping was carried out using the Illumina Infimum HD Human610-Quad BeadChip platform, and after quality control procedures (QC) a total of 527,847 autosomal SNPs were genotyped in 1,775 individuals. The subjects either had a diagnosis of schizophrenia in the Danish Psychiatric Central Register (Mors et al., 2011) or were time-matched controls without schizophrenia records at least until the time of the first diagnosis for the patient. For each SNP, genotypic data was missing for an average of 3-4 subjects (0.19%), and only 82,382 (16%) of the SNPs were completely observed for all subjects. The QC ensured that the maximum proportion of missing genotypes per SNP was 1.0% (18 individuals). Correspondingly, each subject missed on average 0.19% of the SNPs ($\approx$ 1006) and **everyone** had at least 19 SNPs (0.0036%) with unobserved genotypes. Only 268 (15%) of the individuals were missing genotypic data for less than 100 SNPs, whereas the worst case (after QC) missed genotypes for not less than 16,820 SNPs, i.e. a proportion of 3.2%. Thus, if we were to include only persons with completely observed genotypes, none would be left for analyses.

3.3.2 How should the reference panel be chosen?

To impute genotypes for unobserved markers, we need a reference panel of phased haplotypes for which both these, and a reasonably amount of overlapping markers, have been genotyped. The most commonly used reference panels so far have with no doubt been HapMap samples (International HapMap Consortium, 2005), especially HapMap2 (International HapMap Consortium et al., 2007) and HapMap3[14] (International HapMap 3 Consortium, 2010). However, it is likely that the haplotypes from the 1000 Genomes Project (The 1000 Genomes Consortium et al., 2010) will now be used increasingly frequently. For the imputation of the GEMS data (see subsection 3.3.1), we decided to use HapMap3 release 2 as the reference—the main reason being that it was available on the IMPUTE2 web site with positions from the same build (NCBI36) as our data.

HapMap3 contains data from the following population samples: African ancestry in Southwest USA (ASW); Utah residents with Northern and Western European ancestry from the CEPH collection (CEU); Han Chinese in Beijing, China (CHB); Chinese in Metropolitan Denver, Colorado (CHD); Gujarati Indians in Houston, Texas (GIH); Japanese in Tokyo, Japan (JPT); Luhya in Webuye, Kenya (LWK); Mexican ancestry in Los Angeles, California (MEX); Maasai in Kinyawa, Kenya (MKK); Toscans in Italy (TSI); Yoruba in Ibadan, Nigeria (YRI). So should we use the CEU sample because we expect this to match Danish ancestry best? Using older versions of e.g. IMPUTE the answer would probably be yes. But with the implementation of the algorithm by Howie et al. (2011) in current newer versions of IMPUTE2, the answer is to use the full set of haplotypes from all 11 populations and let the program decide which subset of these are optimal to use for each subject.

3.3.3 Issues regarding strand alignment.

Another issue is that the genotypes are expressed relative to the '+' or '-' strand of the human genome reference. Obvious problems arise if e.g. a SNP with alleles A and G relative to the '+' strand in the reference should be used for imputation in a study where this SNP was genotyped relative to the '-' strand, i.e. having T and C alleles. This is handled either by making sure that reference and study samples are relative to the same strand or by submission of a file with strand information to the algorithm (there is an option for doing so in IMPUTE2). A strand file to be used for IMPUTE2 simply contains two columns: first column being the position (from the relevant build) and the second column should contain '+' or '-'. As positions are only unique within chromosome, it is necessary to use separate strand files for each chromosome.

During the imputation of the GEMS data (which was called to the '+' strand) we got some error messages about mismatching alleles which we had to deal with. The reference files should have been aligned to the '+' strand though a few conflicts might be expected (Bryan Howie, personal communication, July 2012). Using a strand file[23] which was generally in agreement with the genotyped alleles, we observed the following differences:

- 140 cases where the strand file suggested '-' strand but with same alleles as in the reference.
- 418 cases where the strand file said '+' but disagreed with the reference on alleles.
- 45,905 SNPs on the chip were not in the reference.
- 1 SNP was not in the strand file but was in the reference.
- 8 SNPs were neither in strand nor in reference.
- 1 SNP had different alleles in the strand file than in our genotypes (and it was noted as 3-4 allelic in databases).

The question was of course if these disagreements were due to errors in the strand files, reference files, study genotypes or a combination. Generally, however, the software was able to decide the strand correctly during imputation except for A/T and G/C SNPs. The problem here being that A is complementary to T and G is complementary to C, i.e. it is never really possible to tell from the genotypes. The usual workaround this seems to be to exclude these SNPs (or include them and accept that there may be errors). Since we also discard a number of SNPs during QC, this is probably not a big issue. Our layman solution to the strand issues was to revise the strand file according to the following:

- '+' strand if study alleles equal reference alleles, and changing '-' to '+' in the strand file if needed.
- '-' strand if study alleles agree with strand file alleles but disagree with reference alleles, i.e. changing '+' to '-' in the strand file if needed.
- Use strand file for SNPs which were not in the reference sample, except for a very few disagreements for which we updated with info from the UCSC[4] Genome Browser.

3.3.4 Speeding up computations.

It is a computer intensive task to impute, so let us close this section with some practical considerations concerning computation time and ways to speed it up. To get an idea of

[23] `http://www.well.ox.ac.uk/~wrayner/strand/`

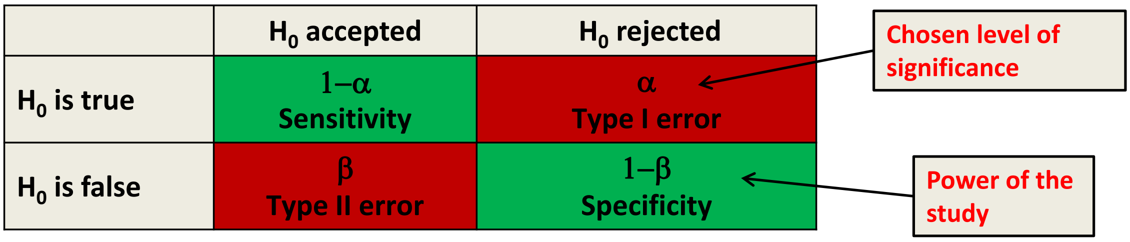
computation time we note that it took almost 2 hours[24] to impute a 5 Mb region (a chunk) on chromosome 22 with 906 genotyped SNPs (851 of which were in the reference) and further 1562 SNPs only in the reference panel. Scaling this up to a whole genome level means that it would take 1-2 month to complete imputation if we processed one chunk at a time. However, using a large cluster of computers this may be done for many chunks simultaneously and taking this to the limit, 35–40 nodes with 16 cores each should be enough to calculate the whole genome for the GEMS data in one go using approximately 2 hours (maybe faster depending on hardware) instead of almost 2 months.

Moreover, IMPUTE2 is well suited for this parallelisation as the accuracy of the algorithm is higher over short regions and actually the recommendations (and default settings) is to use chunk sizes of maximum 5 mega bases (Mb) and a 250 kb buffer zone on each side to prevent edge effects (Howie et al., 2011).

Another way to speed up imputation is to use a so-called pre-phasing step (Howie et al., 2011). The basic idea of this is to ”pre-phase” the study genotypes to produce best-guess haplotypes, and then impute into these estimated haplotypes in a separate program run. In contrast to this, the original IMPUTE2 method integrates over the unknown phase of the study data. Pre-phasing leads to a small loss of accuracy since the estimation uncertainty in the study haplotypes is ignored, but this allows for very fast imputation. This speedup is especially important because modern reference collections (such as those from the 1000 Genomes Project) are frequently updated and expanded, so that many investigators would benefit from ”re-imputing” their datasets following each reference panel update. The pre-phasing step needs to be performed just once per study dataset, so re-imputing is computationally cheap. On the other hand, if a cluster of computers is available for the imputation, then it may not be worth trading accuracy for time.

3.4 Correction for multiple comparisons

Whenever doing statistical testing, it is important to be aware of the risk that conclusions drawn from the tests may be wrong. In principle we operate with two types of errors (type I and type II) as indicated by the confusion matrix in figure 3.1.



	H_0 accepted	H_0 rejected
H_0 is true	$1-\alpha$ Sensitivity	α Type I error
H_0 is false	β Type II error	$1-\beta$ Specificity

Chosen level of significance
Power of the study

Figure 3.1 Confusion matrix. The table shows the two types of errors, type I and type II, considered when testing a null hypothesis. When the significance level (tolerated type I error rate) is chosen, the type II error rate depends on the design (e.g. sample size) and outcome (e.g. effect size).

[24] Microsoft Server 2008, dual-core Intel 64 Family 6 Model 44 Stepping 2 GenuineIntel 1̃594 Mhz, 144 Gb RAM

If m independent tests are made simultaneously under the same null hypothesis (e.g. no association) there will be a probability of

$$\begin{aligned} P(\text{at least one type I error}) &= 1 - P(\text{no type I errors}) \\ &= 1 - P(\text{no type I error in each test})^m \\ &= 1 - (1 - P(\text{type I error in each test}))^m = 1 - (1-\alpha)^m \end{aligned}$$

to make at least one type I error and thus concluding wrongly that an association is significant, i.e. a false positive conclusion. The collection of tests is known as the family of tests and the probability $P(\text{at least one type I error})$ is called the family-wise error rate (FWER) (see e.g. Shaffer, 1995).

The problem is that we choose the significance level $\alpha \in [0,1]$ (typically 0.05) to be the risk we are willing to take of making this false conclusion, but $1-(1-\alpha)^m \geq \alpha$ and it is equal only if $\alpha = 0$, $\alpha = 1$ or $m = 1$. It is easy to see that if $0 < \alpha < 1$ then $1-(1-\alpha)^m$ converges to 1 when m becomes large, see figure 3.2. Under the assumption of independent tests the sum of rejections is binomially distributed $bi(m, p)$ with $p = P(H_0 \text{ rejected})$. Under the null hypothesis, $p = \alpha$ and the mean number of false rejections will be $m \cdot \alpha$ (see figure 3.2).

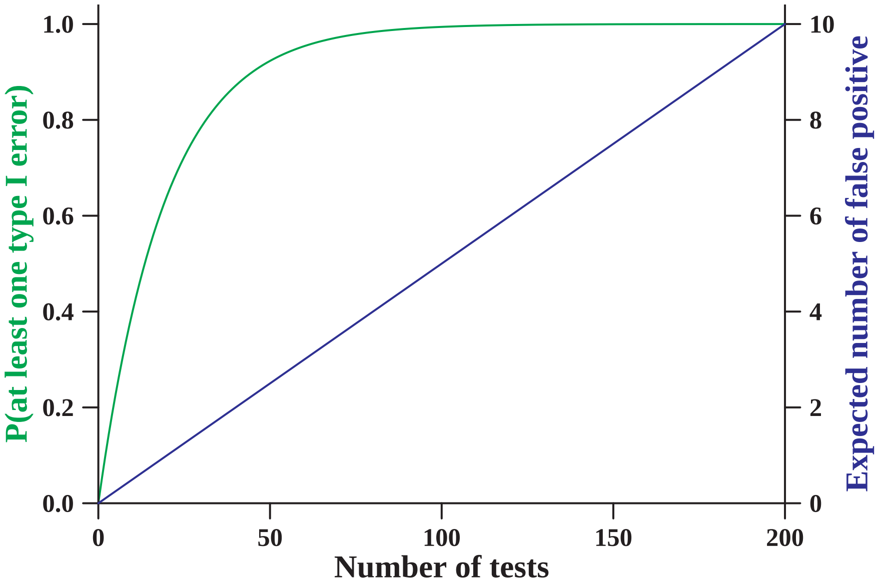


Figure 3.2 Type I errors. The green curve shows the risk of at least one type I error (false rejection) as a function of the number of simultaneous independent tests under the assumption of a true null hypothesis and with a significance level chosen to be 0.05. The blue curve is the corresponding expected number of false rejections (false positive).

Clearly, the risk of at least one false discovery converges very quickly to 1 and the expected number increases linearly. Thus, if we calculate 500,000 single-marker tests from a GWAS and assumes (wrongly) that these tests are independent, we should expect 25,000 false discoveries. The quick and dirty (but easy to remember) trick to control FWER is simply to divide α by the number of tests, i.e. to determine significance at a threshold of α/m. This is known as the (simple) Bonferroni correction and it may be very conservative (too few rejections). Corresponding adjusted p-values are calculated simply as the minimum between 1 and the original p-value multiplied by the number of tests. There are various other versions of Bonferroni corrections (see e.g. Shaffer, 1995) but they are all conservative if the tests are dependent e.g. due to LD between genetic markers. Attempts have been made to compensate for this conservatism by calculation of an ”effective” number of tests (e.g. Nyholt, 2004), i.e. determining a corresponding number of independent tests to control for in Bonferroni-type corrections.

In paper 5 correction for multiple testing was considered by Hommel’s method of controlling the family-wise error rate (Hommel, 1988), which is more powerful (Shaffer, 1995) than simple Bonferroni correction. In paper 1 and 2, to account simultaneously for the nine different SNP and haplotype association tests, permutation adjusted p-values were calculated using a step-down maximum-statistics approach corresponding to the algorithm from Box 2 in Dudoit et al. (2003).

Another method, which has become fairly standard in genetics, is to use the false discovery rate (FDR) that was introduced by Benjamini et al. (1995) and provides a less stringent and potentially more powerful alternative to Bonferroni strategies, by controlling the proportion of wrong rejections of the null hypothesis rather than controlling for no rejections. In short, FDR corrections have greater power at the expense of increased type I error rates (Benjamini et al., 1995; Shaffer, 1995).

Permutation testing is probably the gold standard and would solve the problem attributed to dependency between tests. It may however be computationally infeasible and, furthermore, for interaction analyses it may be difficult to define or decide on an appropriate null hypothesis under which permutations can be done. Obtaining permutation-based p-values for single-marker tests in a case-control design can be done by the following steps:

1. Calculate a test statistic using the observed data, $T_{\text{obs}} \geq 0$ say.
2. Shuffle case-control labels, i.e. draw each subjects affection status without replacement from the pool of labels (case/control).
3. Calculate the test statistic for this permuted data set, T_{perm}.
4. Repeat 2 and 3 a large number of times, B.
5. Calculate the permutation-based p-value P_{perm} as

$$P_{\text{perm}} = \frac{1}{B+1} \sum_{b=0}^{B} \mathbb{1}\{T_{\text{perm},b} \geq T_{\text{obs}}\},$$

   where $\mathbb{1}$ is the indicator function which is one if the statement in the parenthesis is true, and zero otherwise, and $T_{\text{perm},0} = T_{\text{obs}}$.

3.5 The Landscape Method (paper 6)

A motivating example for the *Landscape* method of summarising sequentially ordered tests (or, generally, stochastic variables) was given in subsection 1.1.7 and figure 1.1. This example corresponds to the first 20 positions of the example in Section 2.1 and Figure 1 of paper 6. Paper 6 is rather technical and the results are general, with the intention to be useful for other fields than SNP-based association studies. In the present section we will convey the ideas and measures by going through the results for the motivating example. The more technical details can be found in paper 6 with proofs of theorems, propositions and other theoretical claims in the Appendix. Details on a Python implementation of the method are expected to be available in the submitted version of the paper together with simulation results. An R implementation (in terms of scripts), that can be used for most parts of the method, is shown in appendix A.2. These R scripts were used both for the real data example (see section 2.1.7 and Section 4 in paper 6) and for the calculations shown in the present section.

3.5.1 The landscape and its segments

We consider a consecutive sequence of random variables $Z_1, Z_2, \ldots, Z_K$. The values of Z_k is shown above or under the points in figure 1.1 and in the extended version in figure 3.3 which include the segments defined below. The path or landscape is the accumulated sum

$$A_k = \max\{0, Z_k + A_{k-1}\}, \quad k = 1, 2 \ldots K,$$

with $A_0 = 0$ (see (2.4) in paper 6). Furthermore we set $A_{K+1} = 0$ to enable some later definitions—but A_0 and A_{K+1} are not marked in the plot. Note that in the paper, $\{1, 2, \ldots, K\}$ is defined more generally as a finite or infinite *index set* $\mathbb{K}$ but we will only consider finite sets here. Nevertheless, it is handy to refer to $\{1, 2, \ldots, K\}$ by $\mathbb{K}$ and we will do so.

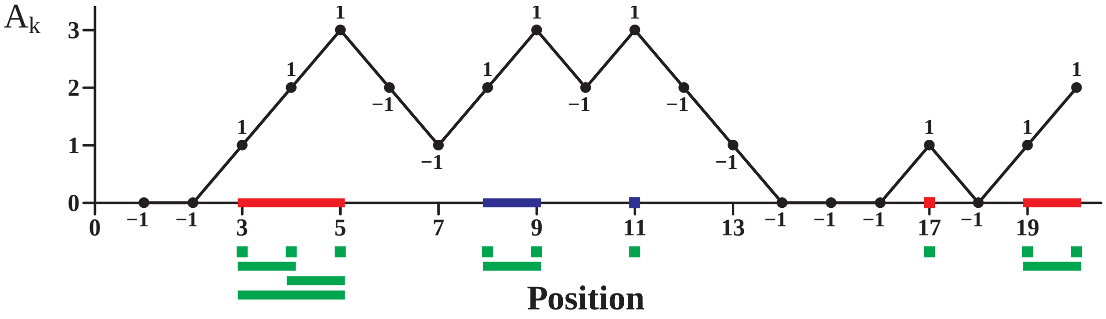


Figure 3.3 Landscape and maximal segments for the motivating example. The coloured bars indicate: independent segments (red), dependent segments (blue), and all segments (green). The union of independent and dependent segments constitutes the maximal segments of the sequence.

Let U_{nm} be partial sums of Z_k's (see (2.1) in paper 6):

$$U_{nm} = \sum_{k=n}^{m} Z_k, \quad 1 \leq n \leq m \in \mathbb{K}.$$

To give just a few values from the motivating example: $U_{3,5} = 3$, $U_{9,12} = 1$ and $U_{13,13} = -1$.

We define a *segment* (Definition 2.2 of paper 6) to be a closed interval $[n, m]$ for which $U_{nk} > 0$ and $U_{km} > 0$ for all values of k in the interval. An examples of a segment in figure 3.3 is $[3, 5]$ because $U_{3,3} = 1 > 0$, $U_{3,4} = 2 > 0$, $U_{3,5} = 3 > 0$, $U_{4,5} = 2 > 0$ and $U_{5,5} = 1 > 0$.

In the special case where, as in the motivating example, $Z_k \in \{-1, 1\}$, it follows that for segments $[n, m]$ of length $m - n + 1 \leq 4$ all corresponding variables Z_k must be 1. To see this observe first that for all segments $[n, m]$ the definition requires $Z_n > 0$ and $Z_m > 0$. Secondly, as also $Z_n + Z_{n+1} > 0$ and $Z_{m-1} + Z_m > 0$ and $Z_k \in \{-1, 1\}$ we see that $Z_{n+1} = 1$ and $Z_{m-1} = 1$. But then $U_{ij} > 0$ for all $i \in [n, m]$ and all $j \in [n, m]$ and the requirements of the definition (Definition 2.2 of paper 6) are fulfilled. In the motivating example this means (and is easy to check) that $[3, 3]$, $[3, 4]$, $[4, 4]$, $[4, 5]$ and $[5, 5]$ are also segments. If $m - n + 1 > 4$ we also see that Z_{n+2} and Z_{m-2}

are the first variables in the sequence $Z_n, \ldots, Z_m$ that may be -1. Note that for general variables Z_k only Z_n and Z_m are also guaranteed to be segments, see also Figure 1 of paper 6.

Given a segment, it seems natural to ask if it is possible to widen it without destroying the requirements for it to be a segment. Thus we define a *maximal segment* to be a segment which is not contained by another segment, i.e. a segment that cannot be enlarged. In the example just given $[3,5]$ is such a maximal segment because $U_{2,2} = Z_2 = -1 < 0$ and $U_{6,6} = Z_6 = -1 < 0$ so if we extend the interval to either of the sides, it will no longer be a segment according to Definition 2.2. In figure 3.3 the red and blue bars show all maximal segments of the sequence. In subsection 3.5.2 we give a recursive algorithm that can be used to find all maximal segments of the sequence. The partial sum U_{nm} is used as the *score of a maximal segment* $[n,m]$, i.e. the score of $[3,5]$ is 3.

3.5.2 Recursive algorithm for finding maximal segments

Equation (2.5)-(2.7) in Section 2.3 of paper 6 presents an algorithm to find all maximal segments of the sequence. The segments found by this algorithm are disjoint (non-overlapping and non-adjacent) and we show (Proposition 2.9 in paper 6) that the maximal segments are precisely the segments found by the algorithm.

We start by defining a *section* to be an interval $S_i = [s_{i0}, t_{i0}]$ between a start point $s_{i0} > t_{i-1,0}$ and a termination point $t_{i0} \geq s_{i0}$ $(i = 1, 2, ..., I)$, with $t_{00} = 0$. Here starting points s_{i0} are defined as the first time $Z_k > 0$ (and thus $A_k > 0$) after last termination point $t_{i-1,0}$, and termination points are defined as the last point after s_{i0} where $A_k > 0$, i.e. $A_{t_{i0}+1} = 0$ (see (2.5) in paper 6). Thus $A_k > 0$, for all $k \in S_i$, i.e. a section is an interval where the landscape is above sea level so to speak. In the motivating example $s_{1,0} = 3$ and $t_{1,0} = 13$, i.e. $[3,13]$ is a section. Continuing, $s_{2,0} = 17$, $t_{2,0} = 17$, $s_{3,0} = 19$ and $t_{3,0} = 20$ (since $A_{21} = 0$ by definition), i.e. $[17,17]$ and $[19,20]$ are also sections.

For completeness we should note that if $Z_k > 0$ for at least one k then there is at least one section, at least one segment, and thus at least one maximal segment, otherwise there are none. It is therefore natural to require that Z_k can take both positive and negative values. In the motivating example this requirement is fulfilled as $Z_k \in \{-1, 1\}$

In connection to the sections we define a maximum score $Y_{i0} = \max\{A_k \mid k \in S_i\}$ in each section and an index $e_{i0} = \min\{k \in S_i \mid A_k = Y_{i0}\}$ of the first time this maximum is obtained (see (2.6) in paper 6). Returning to the motivating example: the maximum score of $S_1 = [3,13]$ is $Y_{1,0} = 3$ obtained first time for $e_{1,0} = 5$; maximum score $Y_{2,0} = 1$ of $S_2 = [17,17]$ is obviously only obtained once and $e_{2,0} = 17$; and the maximum score $Y_{3,0} = 2$ for $S_3 = [19,20]$ is obtain first and only time in $e_{3,0} = 20$.

The values s_{i0}, t_{i0}, Y_{i0} and e_{i0} initiates the recursion for $j > 0$ given in (2.7) of paper 6. Note that the recursion is run separately for each section $i = 1, \ldots, I$, i.e. i is a fixed number in this recursion:

$$\begin{aligned}
s_{ij} &= \min\{k \in S_i \mid k > e_{i,j-1}, A_k > A_{k-1}\}, \\
t_{ij} &= \min\{k \in S_i \mid k \geq s_{ij}, A_{s_{ij}-1} \geq A_{k+1}\}, \\
Y_{ij} &= \max\{A_k \mid k \in [s_{ij}, t_{ij}]\}, \\
e_{ij} &= \min\{k \in [s_{ij}, t_{ij}] \mid A_k = Y_{ij}\}.
\end{aligned}$$

The recursion stops the first time s_{ij} is not defined, i.e. when either $e_{i,j-1} = t_{i0}$ or $s_{ij} = \emptyset$ because $A_k \leq 0$ for all $k > e_{i,j-1}$. That is, the recursion stops when we have traversed all elevated parts of the landscape within the section. From these we get the intervals $[s_{ij}, e_{ij}]$ which can be seen to be segments according to Definition 2.2—see the argument just before Definition 2.8 in paper 6.

Calculating these for S_1 in the motivating example we find: $s_{1,1} = 8$, $t_{1,1} = 13$, $Y_{1,1} = 3$, $e_{1,1} = 9$; $s_{1,2} = 11$, $t_{1,2} = 12$, $Y_{1,2} = 3$, $e_{1,2} = 11$. That is, we find three intervals in S_1: $[3,5]$, $[8,9]$ and $[11,11]$. For $i = 2,3$ we see that $e_{i0} = t_{i0}$ so there are no further intervals than $[17,17]$ respectively $[19,20]$ to be found.

Now the first segment of S_i is called the *independent segment* of this section (Definition 2.8 of paper 6) and remaining segments of S_i are called *dependent segments*. In the motivating example $[3,5]$, $[17,17]$ and $[19,20]$ are independent segments whereas $[8,9]$ and $[11,11]$ are dependent segments.

We have already noted that $[3,5]$ is a maximal segment with score $Y_{1,0} = 3$. That this holds in general for the intervals found by the algorithm is the content of Proposition 2.9 in paper 6. This proposition furthermore states that the scores of the dependent segments depend on the score of the independent segment of S_i, in the sense that these scores are not larger than the score for the corresponding independent segment. This also holds for the motivating example, see the numbers above. In figure 3.3 (and Figure 1 of paper 6) we have shown all segments (green bars) below the x-axis, and the independent (red bars) and dependent (blue bars) maximal segments are indicated on the x-axis.

3.5.3 Evaluation of the signal

The approximate distributional properties of the scores Y_{i0} are shown in Section 3.1 of paper 6 in the case where Z_k is a sequence of independent random variables. To proof the results, aggregation in the opposite direction (from 20 to 1 in the motivating example) is needed. These results are indicated in Section 2.4 of paper 6, and the landscape and maximal segments obtained by this can be seen in Figure 2 of paper 6. One may notice by comparing Figure 1 and Figure 2 of paper 6 that the maximal segments are the same but which of them are independent and dependent vary with direction. Nevertheless, since we do not expect the independence assumption to be valid when we consider tests of genetic markers, we will not go into the case of independent variables here.

The situation with non-independent random variables is treated in Section 3.2 of paper 6 where two approaches are given. Approach 1 requires some homogeneity in distribution across the sequence which may be obtained if $(Z_1, \ldots, Z_{|K|})$ forms a stationary sequence, i.e. if the distribution of any sub-sequence $(Z_k, Z_{k+1}, \ldots, Z_{k+j})$ does not depend on the location k. However, this assumption is unlikely to be true for most collections of genetic markers.

In the general case (Approach 2 in Section 3.2 of paper 6) we let $Y(k) = Y_{ij}$ for $k \in [s_{ij}, e_{ij}]$ (i.e. for k in a maximal segment, see subsection 3.5.2) and $Y(k) = 0$ otherwise (k outside maximal segments). Let us consider the motivating example to grasp this. Here $Y(k)$ is 0 for $k \in \{1,2,6,7,10,12,13,14,15,16,18\}$, 1 for $k = 17$, 2 for $k \in \{19,20\}$, and 3 for $k \in \{3,4,5,8,9,11\}$.

Now to evaluate the significance of the scores we apply a bootstrap procedure to get B bootstrapped samples of the data $(Z_1^b, \ldots, Z_{|K|}^b)$, $b = 1, \ldots, B$. This procedure will generally depend on the data but in the usual setting of association studies (case-control data), this may be done by shuffling the affection (case/control) status. For each bootstrap sample we find the maximal segments using the recursion (subsection 3.5.2) and obtain $Y^b(k)$ as just described for each position in each sample. We then obtain the following approximation of the distribution for

$Y(k)$ (see (3.13) in paper 6):

$$P(Y(k) \geq y) \approx \frac{1}{B+1} \sum_{b=0}^{B} 1(Y^b(k) \geq y),$$

where $b = 0$ denotes the original sample. That is, to obtain a p-value for position k we simple calculate the proportion of bootstrap samples for which the bootstrap score $Y^b(k)$ is at least as extreme as the observed score.

This makes sense since if $\{Y^1(k) \geq y\}, \dots, \{Y^B(k) \geq y\}$ is a series of independent identically distributed events then $\sum_{b=0}^{B} 1(Y^b(k) \geq y) \sim bi(B+1, p)$ and thus $E[\frac{1}{B+1} \sum_{b=0}^{B} 1(Y^b(k) \geq y)] = p = P(Y^b(k) \geq y)$. Therefore $\frac{1}{B+1} \sum_{b=0}^{B} 1(Y^b(k) \geq y)$ is an unbiased estimator of p and will by the law of large numbers approximate this probability. From the central limit theorem it follows that $\frac{1}{B+1} \sum_{b=0}^{B} 1(Y^b(k) \geq y) \approx N(p, \frac{p(1-p)}{B+1})$. So the precision increases (standard deviation decreases) in order of $1/\sqrt{B+1}$. The *only* problem remaining is then to draw the samples such that $Y^b(k)$ has same distribution as $Y(k)$.

The p-value obtained by this bootstrap procedure is a value that indicates the significance of the point being in a maximal segment. We notice that bootstrap p-values equals 1 for all k outside maximal segments of the observed sample, since for these points $Y(k) = 0 \leq Y^b(k)$ for all $b = 1, \dots B$. So the results are only non-trivial for positions inside these maximal segments.

For the motivating example, we can simulate new samples of same size as the original data under the null by simply drawing from $\{-1, 1\}$ with replacement. Alternatively, we can shuffle the signs of Z_k's which corresponds to drawing samples from $\{Z_1, \dots, Z_K\}$ without replacement. As a third alternative, we can do real bootstrapping where we again sample from $\{Z_1, \dots, Z_K\}$ but now with replacement (Davison et al., 1997). To get an idea of the difference and accuracy of the approximation we ran the bootstrap procedure 100 times for each of these three variants using $B = 999$ bootstrap samples in each run. The results for positions in maximal segments are shown in table 3.3

Finally, we may calculate a Bonferroni corrected threshold of significance for multiple testing by dividing the significance level α with the mean number of maximal segments as approximated by the average number of maximal segments in the permutation-samples. We find the following average (and confidence limits) over the 100 repetitions of the mean number of maximal segments: $_{4.893}\,4.903\,_{4.913}$, $_{4.406}\,4.415\,_{4.425}$, $_{4.175}\,4.185\,_{4.195}$ for the simulated, sign shuffled and bootstrapped version, respectively. So the corresponding thresholds for significance would be 0.0102, 0.0113 and 0.0119. This should be compared to the threshold of 0.0025 that would result from correction for 20 tests.

Concerning the different ways of making bootstrap samples it appears that simulating under the null, results in the lowest p-values but to be judged also at a slightly lower threshold. The most obvious difference is that the shuffling of signs retains the number of negative and positive values, whereas the simulation procedure on average will draw equally many -1 and 1. The bootstrapping procedure will on average maintain the proportion of -1 and 1 but the proportion will vary from sample to sample.

3.6 Machine learning methods

Machine learning methods are algorithms that are able to improve their performance of certain tasks by use of the available data (see e.g. Meyfroidt et al., 2009). In one of the earliest studies

Table 3.3

Bootstrap p-values for the motivating example. The average with 95% confidence limits calculated by the normal approximation (${}_{L95}\,mean\,{}_{U95}$) of 100 repetitions of permutation-based p-values for the motivating example by use of three different bootstrap procedures: simulating a random sequence of $\{-1,1\}$ values (Simulate); random sampling without replacement corresponding to shuffling signs (Shuffle); random sampling with replacement, i.e. ordinary bootstrapping (Bootstrap). The results are only shown for positions in maximal segments ($[3,5]$, $[8,9]$, $[11,11]$, $[17,17]$ and $[19,20]$) as permutation-based p-values are otherwise fixed at 1.

Position	Simulate	Shuffle	Bootstrap
3	${}_{0.22}\,0.223\,{}_{0.225}$	${}_{0.252}\,0.255\,{}_{0.258}$	${}_{0.334}\,0.336\,{}_{0.339}$
4	${}_{0.237}\,0.24\,{}_{0.243}$	${}_{0.276}\,0.279\,{}_{0.282}$	${}_{0.364}\,0.366\,{}_{0.369}$
5	${}_{0.267}\,0.27\,{}_{0.273}$	${}_{0.305}\,0.308\,{}_{0.311}$	${}_{0.403}\,0.406\,{}_{0.409}$
8	${}_{0.295}\,0.298\,{}_{0.302}$	${}_{0.343}\,0.346\,{}_{0.35}$	${}_{0.452}\,0.455\,{}_{0.458}$
9	${}_{0.306}\,0.309\,{}_{0.312}$	${}_{0.354}\,0.357\,{}_{0.361}$	${}_{0.467}\,0.47\,{}_{0.473}$
11	${}_{0.317}\,0.319\,{}_{0.322}$	${}_{0.366}\,0.37\,{}_{0.373}$	${}_{0.48}\,0.483\,{}_{0.486}$
17	${}_{0.473}\,0.476\,{}_{0.479}$	${}_{0.486}\,0.488\,{}_{0.491}$	${}_{0.548}\,0.552\,{}_{0.555}$
19	${}_{0.377}\,0.38\,{}_{0.384}$	${}_{0.38}\,0.383\,{}_{0.386}$	${}_{0.45}\,0.453\,{}_{0.457}$
20	${}_{0.297}\,0.3\,{}_{0.303}$	${}_{0.328}\,0.331\,{}_{0.334}$	${}_{0.408}\,0.411\,{}_{0.414}$

of the principles of machine learning, Samuel (1959) use the game of checker to investigate how to program a digital computer such that it behaves in a way that seemingly correspond to the process of learning seen in living organisms, e.g. humans. A definition of machine learning from this paper could be: A technique of programming computers to learn from experience and thereby eliminating much of the need for more detailed programming.

Machine learning may be seen as a subfield of the larger branch of computer science known as *artificial intelligence*, i.e. concerning development of systems and software that are able to learn from experience. It is to some extent confused with the term *data mining*, possibly because of a great overlap of the methods used. However, where data mining targets on the discovery of new properties, machine learning is often more devoted to prediction on basis of properties learned from some training data.

Machine learning is usually categorised as *supervised learning* typically used for prediction, and *unsupervised learning* used for descriptive tasks where the target is unknown and the goal is to describe regularities or structure of complex data, and e.g. cluster subgroups that are similar in some perspective. Many different types of algorithms exists: decision trees (classification and regression trees), random forests, artificial neural networks, genetic algorithms and programming, Bayesian networks, support vector machines, Gaussian processes, fuzzy logic etc.

In the study presented in paper 3, we used a version of the popular machine learning and data mining method MDR, model-based MDR (MB-MDR) (Calle et al., 2008), and a version of the machine learning method logic regression (Ruczinski, 2000; Kooperberg et al., 2001; Ruczinski et al., 2003), logic feature selection (logicFS) (Schwender et al., 2011b).

3.6.1 MB-MDR

The multifactor dimensionality reduction (MDR) method was introduced by Ritchie et al. (2001) to improve identification of higher order interaction between genetic markers and/or other factors in disease association analysis. The method is nonparametric and do not make assumptions on the genetic penetrance model, i.e. it is model-free. The basic idea of MDR is to pool multilocus genotypes into two groups: a high-risk and a low-risk group. This way the dimensionality is reduced from the number of factors to just one dimension, the new high/low factor. In the original method the ability of this new factor to classify and predict disease status is evaluated by cross-validation and permutation testing. For an overview of this procedure and a review of the classical version of MDR we refer to Motsinger et al. (2006). Updated Java-based software for this method exists[25] and the method have also been implemented in R (Winham et al., 2011) as the package MDR.

As noted above, we decided to use the MB-MDR approach proposed by Calle et al. (2008) which was found to generally have higher power than MDR, especially in situations with presence of genetic heterogeneity and phenocopies where MDR tends to have less success (Calle et al., 2008; Cattaert et al., 2011). Some of the limitations of the classical MDR are listed in Calle et al. (2008) and handled by the MB-MDR method which furthermore should be computationally more efficient. The improvements include the possibility to adjust for main effects. MB-MDR was first implemented and used for case-control studies, i.e. binary traits, but later extended to quantitative traits (Mahachie John et al., 2011) and censored traits (Van Lishout et al., 2013). The principal difference between MB-MDR and classical MDR is that MB-MDR only merges genotype combinations that show significant evidence of high or low risk. The remainder, i.e. combinations with no evidence or insufficient sample size, are merged into a third category. MB-MDR thus pool multilocus genotypes into three groups rather than two groups as in MDR. The idea is to avoid noise from combinations (cells) that are not important for the association effect either due to power issues from small counts or low effect size, or because the null is true.

The procedure of MB-MDR consists principally of three steps, see Figure 1 of Cattaert et al. (2011). In step 1, all possible combinations of the k factors ($k = 1, 2, \ldots$) are represented in the k-dimensional (k-D) space and each cell is tested for association with the trait. The choice of test depends on the trait type and may as such also be parametric or nonparametric. In step 2, the p-values for the test statistic calculated in step 1 are thresholded against some reference critical value p_c which is set to 0.1 per default in the available software (see below) as recommended by Cattaert et al. (2011) and also applied in Calle et al. (2008). In essence, for a binary trait, a threshold of 1 resembles the classical MDR, c.f. Cattaert et al. (2011). Cells with $p < p_c$ are then classified as high risk (H) or low risk (L) depending on the direction of the effect, whereas cells with $p \geq p_c$ are classified as no risk evidence (O). In practice (i.e. in the software) a second threshold (the '--m' option with 10 as default) defines a minimum group size (cases+controls for a disease trait) for which it is statistically relevant to calculate a test statistics and a p-value. Cells with a group size less than this second threshold is classified as O. As in classical MDR, we obtain a considerable reduction of dimensionality by use of this new one-dimensional three-levelled categorical variable. To give an example, if we search for 3-way interactions between SNPs then for each combination of 3 SNPs, we would go from a 3-D 27-celled cubic matrix problem to the much simpler 1-D vector of length 3. Now (still in step 2) a second round of association tests is calculated for these 'HLO' vectors and again the method allows for different testing strategies (Calle et al., 2008; Cattaert et al., 2011). In step 3, the significance of the test statistics from step 2 are determined with correction for multiple correlated tests. In Calle et al. (2008) the statistic

[25] `http://sourceforge.net/projects/mdr/`

used is a Wald test which is assessed in permutation-based null distributions. In Cattaert et al. (2011) resampling-based step-down *maxT* adjusted p-values (Westfall et al., 1993) are calculated where the statistic (for a binary trait) is the maximum of 1, 2 or 3 one degree-of-freedom χ^2-test statistics calculated for association with the H category versus L and/or association with the H category versus L/O (L and O pooled) respectively association with L versus H/O.

The MB-MDR method have been implemented in R (Calle et al., 2010), the package mbmdr, but we preferred the C++ implementation (Cattaert et al., 2011) which now includes the efficient implementation of the multiple testing algorithm *MAXT* by Van Lishout et al. (2013). A faster algorithm (*speedMAXT*) has been available since version 4.0.1 of the software[26]. The *speedMAXT* algorithm have not been published yet but the idea is to trade a slightly higher false-positive rate for time (personal communication, François van Lishout, January 2014). The statistic currently used by the programme is the maximum of the two tests H vs L/O and L vs H/O—in the earliest version (2.7.5) available from the web page it was possible also to choose the H vs L technique which is the method used in the mbmdr R package. A couple of other variants for multiple test adjustment are available for the 2-D case, e.g. the classical *minP* step-down permutation algorithm (Westfall et al., 1993). During our trials using the software, we encountered a few minor bugs which we communicated to the main author of the software F. van Lishout. These have all been solved in the newest version (4.1.0). The software includes the useful and efficient opportunity to run parallel workflow both for the *MAXT* algorithm and to an even greater extent for the speedMAXT algorithm. It currently handles interactions up to 3-D (k-D, $k \in \{1,2,3\}$), i.e. single-markers, 2- or 3-way interactions, and can be used both for binary, continuous and time-to-event (censored) traits. The mbmdr R package does not have a limit as such on the dimension but we guess that things like memory limitations will have an impact on this—depending also on the number of SNPs to be searched. Using the parallelised workflow of *MAXT*, Van Lishout et al. (2013) were able to search for all SNP-SNP interactions (2-D) in 100,000 SNPs and 1000 individuals within reasonably time, and with the implementation of speedMAXT there may be hope for full scale genome-wide interaction studies to be reachable. An option ('--f') makes it possible to run for example all 3-way interactions between 1 (fixed) environmental factor and all possible SNP-SNP combinations. Obviously, this is much faster than searching also through all possible 3-way interactions between the SNPs and the multiple comparison issue will be (many) orders of magnitude smaller. To exemplify: with 100 SNPs and 1 environmental factor there would be 166,650 3-way interactions but only 4,950 includes the environmental factor. For the adjustment of main effects it is possible to choose co-dominant or additive coding of the genotypes (or no adjustment). Mahachie John et al. (2012) recommends to always account for main effects of the SNPs under investigation for interaction and to do this as an integrated part of using MB-MDR as this adequately controls false positive findings. Furthermore they concluded that the co-dominant correction should be preferred as the additive coding may be insufficient and lead to overly optimistic results (c.f. Mahachie John et al., 2012).

3.6.2 Logic regression

Logic regression is a machine learning method that can be used to detect and quantify importance of genetic interactions in case-control studies. Originally, the methodology was developed to allow inclusion of combinations of binary predictors by Boolean (logic) statements (see appendix A.1.3) to enhance prediction of a response (Ruczinski, 2000), see also appendix A.1.1 for some details on the origin of the method. This was deployed in a regression framework (see appendix A.1.4) and thus the name *logic regression*. The framework includes e.g. linear regression, logistic

[26] http://www.statgen.ulg.ac.be/software.html

regression and Cox regression but may be any type of regression as long as a scoring function can be defined. The search over the entire space of combinations then aims at optimising this scoring function. It is a prerequisite that the predictors are either binary (0/1, yes/no etc.) or can be formulated as a Boolean combination of binary variables. Thus, continuous variables can only be used after discretisation, though they may also enter as covariates in the regression part of the method but without entering the search algorithm as such. Apart from using logic regression in a regression setting it may also be used for classification (or miss-classification). There is a difference in complexity as classification will just search for a single Boolean expression (Boolean combination of binary predictors) for prediction of a binary response whereas in the regression framework, several Boolean expressions may enter and the response need not be binary.

With k binary predictors there are potentially 2^k different combinations each of which can predict a zero or one (control or case, say), that is in principle 2^{2^k} possible *prediction scenarios* (see Ruczinski, 2000). Thus, the number of scenarios grows with double exponential speed and becomes incredibly large even for a relatively small number of predictors: 1 predictor leads to 4 scenarios, for 2 predictors there are 16, for 3 the number is 256, at 4 predictors the number has grown to no less than 65,536 and already at 5 predictors there are billions of combinations. Therefore, there may well for each of l subjects exist two logic trees unique for that individual that predict zero and one, respectively. That is, in total up to 2^l different logic trees might be consistent with the observed predictors of the sample investigated. Consequently a *simulated annealing* search algorithm is used to find the *best fitting* model, see appendix A.1.5.

A problem of simulated annealing is that the process may end up with a model that overfits the data and tools for model selection are therefore needed and also included in some of the software that implements logic regression. As a measure of model complexity, the *model size* is defined to be total number of leaves in the trees of the model (Ruczinski et al., 2003). There are thus two handles than can be turned to change the model size: number of trees and number of leaves in each tree.

Logic regression has been applied in various settings and there exists a growing number of variants and further developments of logic regression (Schwender et al., 2010), see also appendix A.1. The original method (with some extensions) has been implemented as the R package LogicReg. In the study presented in paper 3, we prepare to use the variant called *Logic Feature Selection* (logicFS) that has been implemented as the R Bioconductor[27] package logicFS, which depends also on LogicReg. Depending on trait, this currently (version 1.32.0, September 12, 2013) handles: classification, logistic regression, linear regression and multinomial logic regression. The program uses either bagging (bootstrap sampling) or subsampling to stabilise the search and furthermore returns a variable importance measure (VIM) which can be used to determine the importance of the interactions found, see the Appendix of paper 3.

In the context of gene interactions, an advantage of logic regression is that these interactions need not be known in advanced and interactions are not restricted to pairwise interactions, e.g. SNP-SNP interactions. Since the effects of SNPs are typically small, the main focus will normally be more on detection (association and feature selection) than on prediction, though the latter is also technically possible. Typically, applications of logic regression in genetics have focused on gene-gene interactions, but principally gene-environment interaction can also be investigated if the environmental measure is dichotomous or captured by a combination of dichotomous variables. Again, discretisation may be a way to investigate continuous environmental exposures.

[27] http://www.bioconductor.org

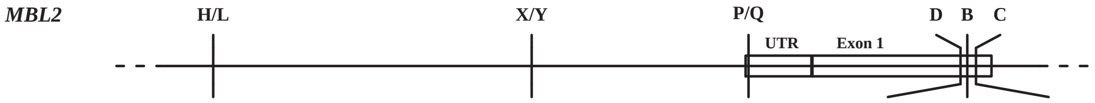
Chapter 4

Results

4.1 Results from paper 1 and 2: MBL and MASP-2

Instead of giving results separately we will here go a step further and show combined tables and figures from the two papers as originally intended. In addition, we will show some results comparing the patient groups with each other (post hoc contrasts) in models where the phenotype has four categories corresponding to controls and patients with schizophrenia, panic disorder and bipolar disorder. Allele frequencies of the genetic markers in *MBL2* are shown in figure 4.1 which is a combination of Figure 1 of paper 1 (Foldager et al., 2012) and the upper part of Table 1[28] in paper 2 (Foldager et al., 2014).

Figure 4.1



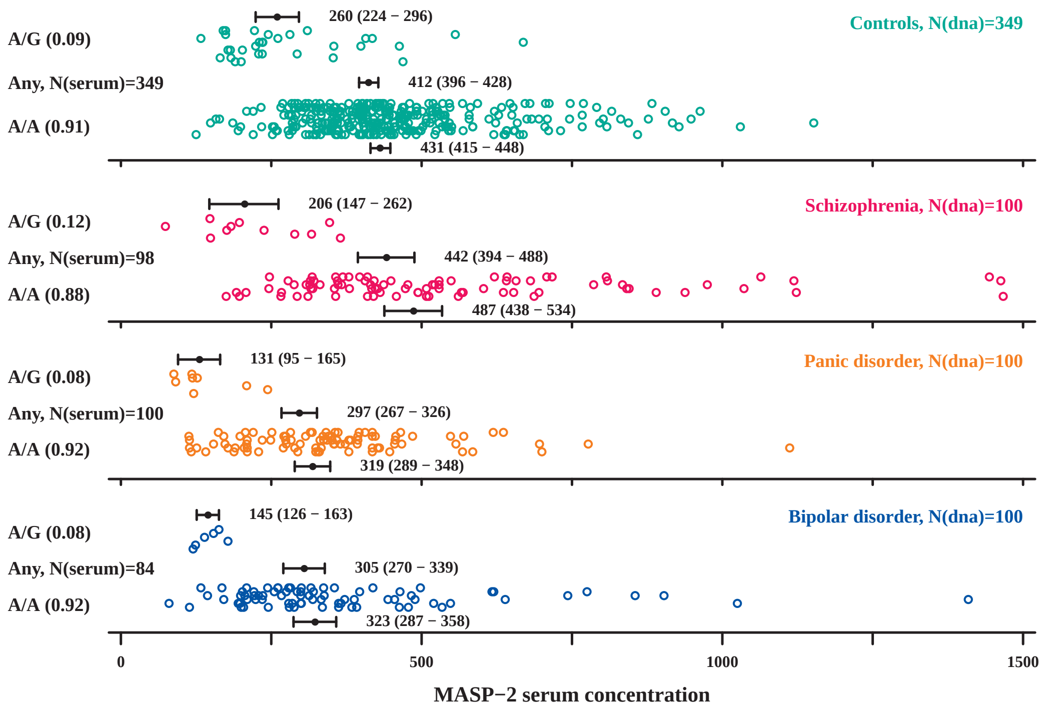
Reference name	**rs11003125**	**rs7096206**	**rs7095891**	**rs5030737**	**rs1800450**	**rs1800451**
Relative position	**− 550**	**− 221**	**4**	**223**	**230**	**239**
Rare allele	**H**	**X**	**Q**	**D**	**B**	**C**
− freq. in controls	**0.39**	**0.20**	**0.21**	**0.07**	**0.14**	**0.01**
− − schizophrenia	**0.28**	**0.27**	**0.20**	**0.05**	**0.19**	**0.03**
− − panic disorder	**0.32**	**0.24**	**0.18**	**0.08**	**0.20**	**0.02**
− − bipolar disorder	**0.36**	**0.29**	**0.20**	**0.07**	**0.12**	**0.02**

Positions and frequencies of genetic markers in *MBL2*. Reference names and positions for the genetic markers in *MBL2* located at 10q21.1. The positions are relative to the untranslated (UTR) start position of exon 1. Allele frequencies of the minor alleles are from 349 controls, 100 patients with schizophrenia, 100 patients with panic disorder, and 100 patients with bipolar disorder.

[28] As an aside note, the first *Total* row in Table 1 of paper 2 is a bit misleading—it is the total number of chromosomes, not the total count of the minor alleles (in the haplotype part these numbers agree).

The allele frequencies of the genetic marker in *MASP2* can be seen in figure 4.2. None were homozygous for the minor G allele which was a bit more frequent in patients with schizophrenia (12%) than in the other groups (8–9%).

Figure 4.2



MASP-2 serum concentration and *MASP2* genotypes. Bootstrapped back-transformed estimates of median MASP-2 serum concentration with 95% CI are presented from two log-gaussian linear models: 1) phenotype only (*any* genotype); 2) interaction between phenotype and A/G genotype status (present/absent). Shown are also the number of subjects in each cohort that was measured (genotypes and serum, respectively) as well as the proportion of subjects carrying the *MASP2* A/G genotype of D105G.

Frequencies of the seven observed haplotypes (see 3.1.5) and the YA/XA/YO two-marker genotypes (see 2.2.1) are shown in table 4.1. From this, a little bit of mental arithmetics reveal that the proportion carrying at least one of the nonsynonymous variants in *MBL2* exon 1 was especially high (50%) for patients with panic disorder. In controls this proportion was 39% whereas it was somewhat lower in patients with bipolar disorder (36%) and higher (compared to controls) in patients with schizophrenia (43%).

4.1.1 Association analysis

The trend test (1 d.f. χ^2) results for *MBL2* and *MASP2* single locus and *MBL2* multilocus markers are shown in table 4.2 (Table 2 in paper 1 and paper 2). For the variant in *MASP2* no significant associations were found though the proportions of D120G A/G heterozygotes were higher in patients with schizophrenia.

Table 4.1

***MBL2* haplotype and multilocus genotype frequencies.** Minor alleles are marked with bold type and the O allele is any of the D, B and C nonsynonymous mutations of exon 1. Multilocus genotypes are grouped with respect to their known association with low, intermediate or normal level of MBL in serum (Olesen et al., 2006).

Counts (prop.)	Controls	Schizophrenia	Panic dis.	Bipolar dis.
***MBL2* haplotype**	N_{hap}=698	N_{hap}=200	N_{hap}=200	N_{hap}=200
HYPA	223 (0.32)	47 (0.23)	50 (0.25)	57 (0.28)
LYPA	35 (0.05)	12 (0.06)	11 (0.06)	8 (0.04)
LY**Q**A	141 (0.20)	35 (0.17)	31 (0.16)	37 (0.18)
L**X**PA	139 (0.20)	53 (0.27)	49 (0.24)	58 (0.29)
HYP**D**	51 (0.07)	10 (0.05)	15 (0.08)	14 (0.07)
LYP**B**	100 (0.14)	38 (0.19)	40 (0.20)	23 (0.12)
LY**QC**	9 (0.01)	5 (0.03)	4 (0.02)	3 (0.02)
Multilocus genotype	N_{geno}=349	N_{geno}=100	N_{geno}=100	N_{geno}=100
Normal MBL level				
YA/YA	119 (0.34)	17 (0.17)	25 (0.25)	29 (0.29)
YA/**X**A	77 (0.22)	36 (0.36)	21 (0.21)	28 (0.28)
Total	196 (0.56)	53 (0.53)	46 (0.46)	57 (0.57)
Intermediate				
XA/**X**A	15 (0.04)	4 (0.04)	4 (0.04)	7 (0.07)
YA/Y**O**	84 (0.24)	24 (0.24)	21 (0.21)	16 (0.16)
Total	99 (0.28)	28 (0.28)	25 (0.25)	23 (0.23)
Low MBL level				
XA/Y**O**	32 (0.09)	9 (0.09)	20 (0.20)	16 (0.16)
Y**O**/Y**O**	22 (0.06)	10 (0.10)	9 (0.09)	4 (0.04)
Total	54 (0.15)	19 (0.19)	29 (0.29)	20 (0.20)

In *MBL2* single-markers, significant association (P=0.006) was found for the H/L marker in schizophrenia and this signal remained when we, post hoc, tried to pool the three groups of patients (P=0.0089). The minor H allele is more frequent in controls and thus showing a protective effect (odds ratios less than one). Nominally significant association with schizophrenia was also observed for the X/Y marker—in this case as a disease predisposing effect of the minor allele.

Table 4.2

***MBL2* and *MASP2* trend test**. Trend tests (1 d.f. χ^2) for association of patient groups with *MBL2* and *MASP2* single locus markers and *MBL2* multilocus genetic markers by use of logistic regressions with an additive effect on a log scale of the minor allele (marked with bold type). The odds ratio (OR) measures the effect of an extra minor allele and OR between the two homozygote variants is therefore this value squared.

P $_{L95}\,OR\,_{U95}$	Schizophrenia	Panic disorder	Bipolar disorder	Pooled cases
MASP2				
D120G (A/**G**)	0.41 0.64 1.35 2.67	0.71 0.36 0.86 1.85	0.71 0.54 1.16 2.78	0.94 0.60 1.02 1.74
MBL2				
Single locus				
H/L (m1)	**0.006** 0.45 0.63 0.88	0.090 0.55 0.76 1.04	0.34 0.62 0.85 1.18	**0.0089** 0.59 0.74 0.93
X/Y (m2)	**0.047** 1.01 1.46 2.12	0.16 0.89 1.31 1.90	**0.0075** 1.14 1.65 2.36	**0.0033** 1.14 1.49 1.95
P/**Q** (m3)	0.65 0.61 0.91 1.34	0.22 0.51 0.78 1.15	0.65 0.61 0.91 1.34	0.30 0.66 0.86 1.14
A/**D** (m4)	0.24 0.32 0.68 1.29	0.93 0.55 1.03 1.84	0.88 0.50 0.95 1.72	0.57 0.57 0.88 1.36
A/**B** (m5)	0.14 0.90 1.34 1.95	0.075 0.96 1.42 2.07	0.33 0.50 0.80 1.24	0.25 0.89 1.18 1.57
A/**C** (m6)	0.24 0.60 1.99 5.90	0.47 0.42 1.57 4.95	0.82 0.26 1.17 4.00	0.31 0.66 1.57 3.91
A/**O** (m7)[a]	0.32 0.84 1.19 1.68	0.067 0.98 1.38 1.95	0.39 0.57 0.85 1.23	0.33 0.88 1.13 1.45
Multilocus[b]				
HYPA	**0.023** 0.47 0.67 0.95	0.066 0.51 0.73 1.02	0.36 0.61 0.86 1.19	**0.015** 0.59 0.74 0.94
LYPA	0.59 0.59 1.20 2.27	0.79 0.53 1.09 2.07	0.56 0.35 0.80 1.62	0.90 0.63 1.03 1.66
LY**Q**A	0.39 0.55 0.84 1.25	0.14 0.47 0.73 1.10	0.59 0.59 0.90 1.33	0.16 0.61 0.82 1.08
YA[c]	**0.011** 0.48 0.66 0.91	**0.0074** 0.48 0.66 0.89	0.14 0.58 0.79 1.08	**0.0013** 0.56 0.70 0.87

[a] The O allele of the A/O marker is any of the D, B and C variants of *MBL2* exon 1.

[b] LXPA, HYPD, LYPB and LYQC are identifiable with m2, m4, m5 and m6, respectively.

[c] XA and YO are identifiable with m2 and m7, respectively.

This effect was significant also in patients with bipolar disorder (P=0.0075) and the significance increased even further by pooling cases (P=0.0033). No other single-markers showed significant association with the disorders though there was a tendency of association with panic disorder for the B variant in exon 1 (P=0.075). The allele frequency for this variant was equally high in patients with schizophrenia and panic disorder (see figure 4.1), and pooling these two groups of patients (not shown in papers or tables) pushed the p-value below the threshold of nominal significance (P=0.038) with an odds ratio of $_{1.02}\,1.38\,_{1.88}$. Low power due to the relatively small sample sizes may well explain why we are not detecting significant associations with the non-silent variants in exon 1.

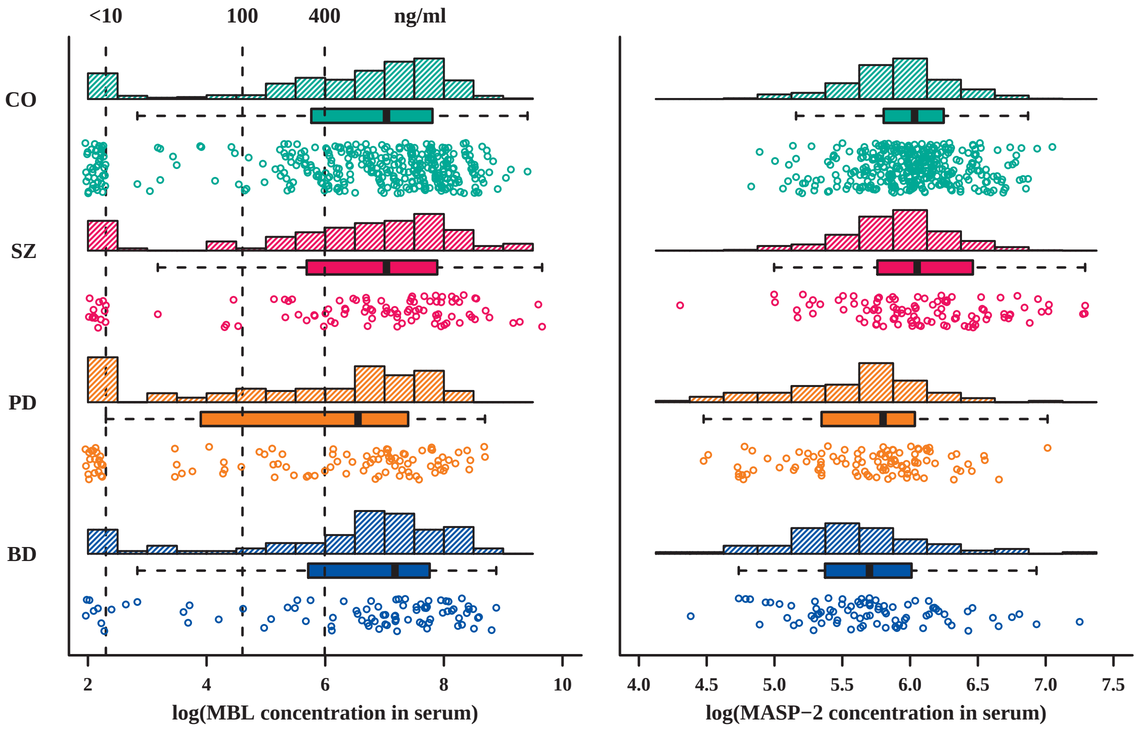
For *MBL2* haplotypes, the HYPA showed nominal significant protective effect against schizophrenia and a tendency (P=0.066) against panic disorder. Pooling all cases, this p-value was 0.015 while pooling just the groups of patients with schizophrenia and panic disorder resulted in a significant p-value of 0.0082 (result not shown in the papers or tables). The tendencies of the effect for the other two haplotypes combining intron variants with the exon 1 wild type A allele (LYPA and LYQA) were also either neutral or protective. This is reflected in the YA two-marker haplotype which in essence is a grouping of these three haplotypes showing a protective effect against schizophrenia (P=0.011) and even more against panic disorder (P=0.0074). Here both the pooling of all cases (P=0.0013) and, even more, pooling schizophrenia with panic disorder groups (P=0.00080) lowered the p-value markedly. The odds ratio for YA association with this last grouping (schizophrenia and panic disorder) was $_{0.51}\,0.66\,_{0.84}$. Furthermore recall that LXPA and XA are equivalent in the present studies and identifiable by the single locus X variant which showed significant associations. Likewise, HYPD, LYPB and LYQC are identifiable with the corresponding single-marker variants in exon 1: D, B and C, respectively.

4.1.2 MBL and MASP-2 serum concentration

In figure 4.3 (Figure 2 in paper 1 and Figure 1 in paper 2) we have shown the distribution of log-transformed MBL and MASP-2 serum concentrations. Note the bulk of measures below MBL detection limit—which were left out when making the histogram of the measured values—and that the MBL distributions are a bit left skewed after log transformation (they were right skewed before transformation). We decided, however, after various model checking (e.g. qq-plots of residuals from linear regressions where the values below detection limit were left out), that using gaussian Tobit regression on the log-transformed data was reasonably enough. For MASP-2 log-transformation took nice care of the right skewness and here there is no detection limit problems to take care of. Thus we analysed MASP-2 with log-gaussian linear regression.

The genetic markers considered were chosen because of their known association with MBL and MASP-2 levels, respectively. In figure 4.4 we show box-and-whiskers plots of log-MBL serum concentration for each haplotype. These clearly indicate the association between serum concentration and the haplotypes of *MBL2* and calculations supported this observation (results not shown). This was expected and not relevant to dwell on too much. Of course, the number of chromosomes (0, 1 or 2) of each haplotype carried by the individual matters for the expression and, to keep it simple (and readable), we show in figure 4.5 the box-and-whiskers plot separately for each two-locus genotype of the YA/XA/YO grouping (see subsection 2.2.1). Here it is very clear that very low MBL levels correlate with carrying two of the exon 1 variants (HYPD, LYPB or LYQC) or one exon 1 variant in combination with the LPXA haplotype. In fact, only one of the carriers of two exon 1 variants had a measurable MBL level and this person (a control with LYPB/HYPD multilocus genotype) still had a very low concentration of just 25 ng/ml. At the other end of the spectrum lies YA (HYPA, LYPA or LYQA) carriers not carrying an exon 1 variant. In between are LXPA homozygotes and the carriers of an exon 1 variant having one of the YA haplotypes on the other strand. The plot looks a bit interesting for the XA/XA subgroup but one should be aware that there are only 30 individuals in the group (15 controls, 4 with schizophrenia, 4 with panic disorder and 7 with bipolar disorder).

Figure 4.3



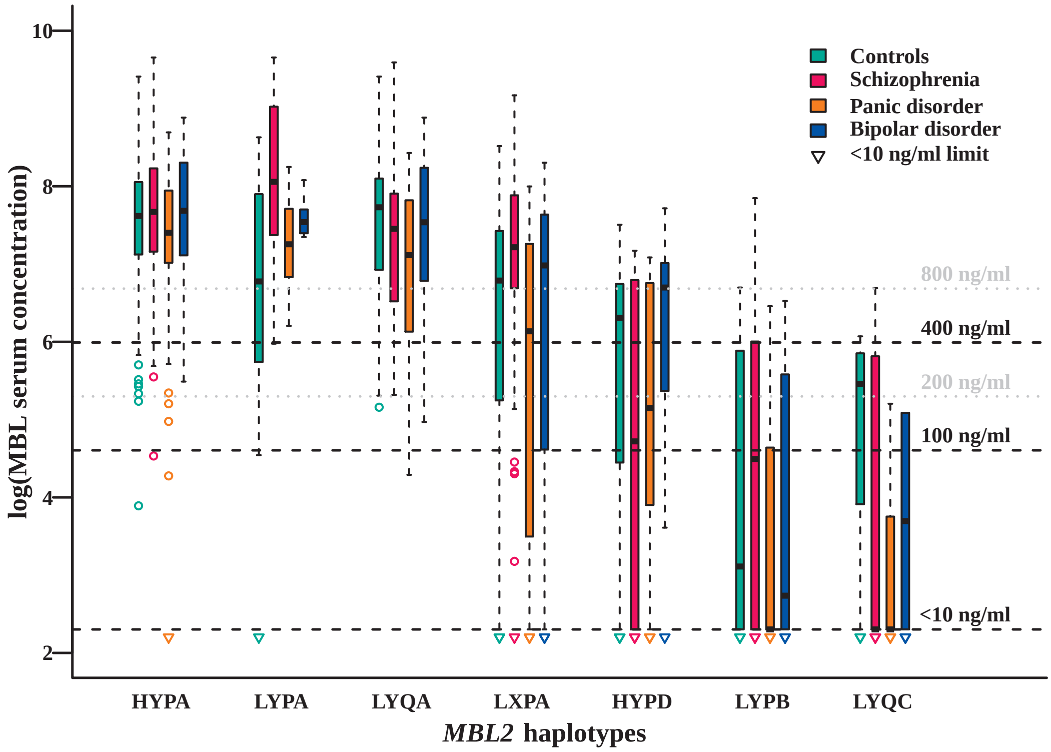
Distribution of MBL and MASP-2 serum concentration. Concentration of MBL and MASP-2 in serum for 349 controls (CO), 98 patients with schizophrenia (SZ), 100 patients with panic disorder without a history of bipolar disorder (PD), and 84 patients with bipolar disorder (BD). Before logarithmic transformation, concentrations were measured in ng protein per ml serum. The vertical lines in the left panel indicate: below MBL detection limit (<10 ng/ml), low MBL level (<100 ng/ml), intermediate MBL level (100–400 ng/ml) and normal MBL level (>400 ng/ml). Histogram, box-plot and scatter plot of the observed concentrations are given for each protein and separately for each group of subjects.

The concentrations divided on the genotypes of the *MASP2* marker are also shown in figure 4.5 and do not indicate association between the *MASP2* marker and MBL levels. The low MASP-2 levels for carriers of the D120G variants G allele of *MASP2* can be seen in figure 4.2.

MBL serum deficiency

The results from categorising MBL levels as *low*, *intermediate* and *normal* (see subsection 2.2.1) are shown in figure 4.6 (Figure 2 in paper 2) and also indicated in figure 4.3. An unusual high proportion of patients with panic disorder (30%) had low MBL level—significantly (P=0.0008) higher than the 15% seen in controls, and with an odds ratio of ${}_{1.4}\,2.4\,{}_{4.0}$. The corresponding proportions for the other phenotypes were 18% and 17% in patients with schizophrenia and bipolar, respectively. Also significantly higher than in controls (P=0.027), 20% of patients with panic disorder had MBL serum concentration below the detection limit (<10 ng/ml), OR: ${}_{1.1}\,1.9\,{}_{3.5}$. This proportion was a bit lower in patients with bipolar disorder (8%) and a bit higher in patients with schizophrenia (13%) than in controls (11%) but within or close to the 10–15% generally seen in populations.

Figure 4.4

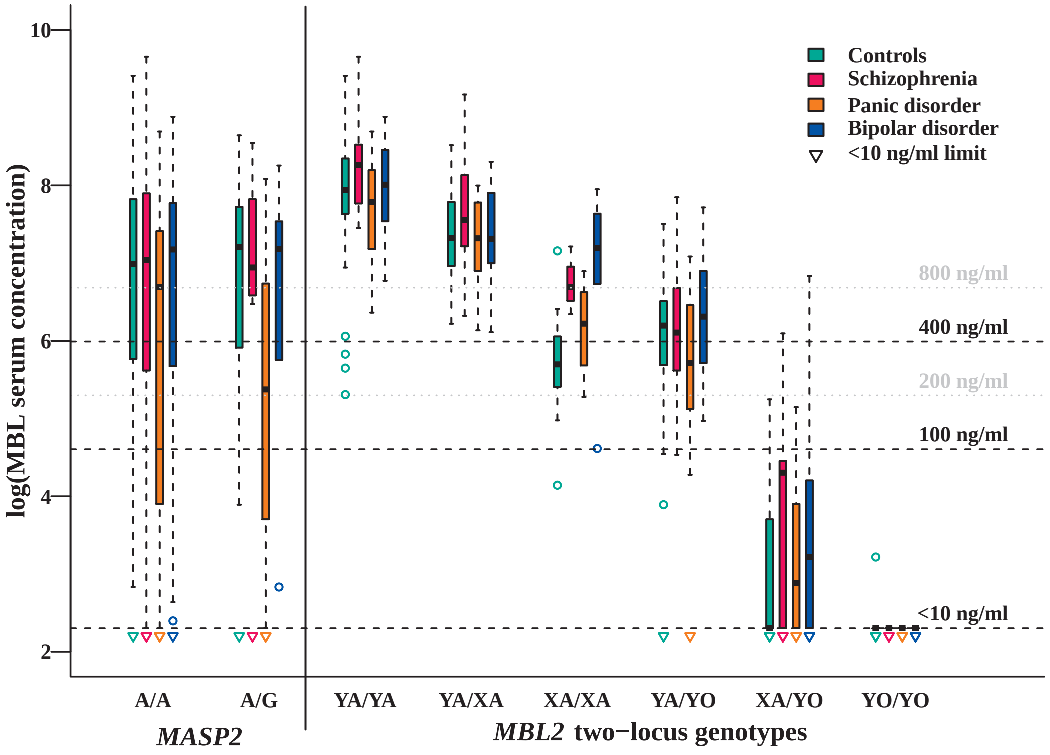


MBL serum concentration on *MBL2* haplotypes. Concentration of MBL in serum for 349 controls, 98 patients with schizophrenia, 100 patients with panic disorder without a history of bipolar disorder, and 84 patients with bipolar disorder. Before logarithmic transformation, concentrations were measured in ng protein per ml serum. The black dotted lines indicate: below MBL detection limit (<10 ng/ml), low MBL level (<100 ng/ml), intermediate MBL level (100–400 ng/ml) and normal MBL level (>400 ng/ml). The gray dotted lines indicate the classification used in Olesen et al. (2006). Box-and-whiskers plots are shown for each of the seven haplotypes and divided on phenotype. Note that each individual carries two haplotypes and thus contributes in duplicate.

MBL Tobit regression

The results from MBL Tobit regressions within the individual patient groups versus controls are shown in Table 3 of paper 1 (Foldager et al., 2012) and Table 3 of paper 2 (Foldager et al., 2014). Disregarding the genetic markers for a minute, we found that patients with panic disorder had significantly lower serum concentration of MBL than controls whereas no association was found with schizophrenia and bipolar disorder. This was also evident from a comparison of quantiles, see table 4.3 (partly Supplementary Table S3 of paper 2). This association is actually also visual both in figure 4.3 and not least in figure 4.6. It stems both from relatively many low values (<100 ng/ml) and from a high proportion of MBL measures below detection limit (<10 ng/ml) for patients with panic disorder. Some explanation likely lies in the larger proportion of patients with panic disorder carrying nonsynonymous exon 1 variants (see figure 4.1 and table 4.1). However, even after adjusting for *MBL2* haplotypes there still were nominal significantly lower MBL levels in panic disorder. Interestingly, the level of MBL was higher in patients with schizophrenia than in controls when we adjusted for the effect of genetic variation in *MBL2*, and this difference was highly significant with a p-value as low as 1.7e-6. From Table 3 of paper 1 it also follows that using a more detailed genotype grouping than those induced by the YA/XA/YO two-marker

Figure 4.5

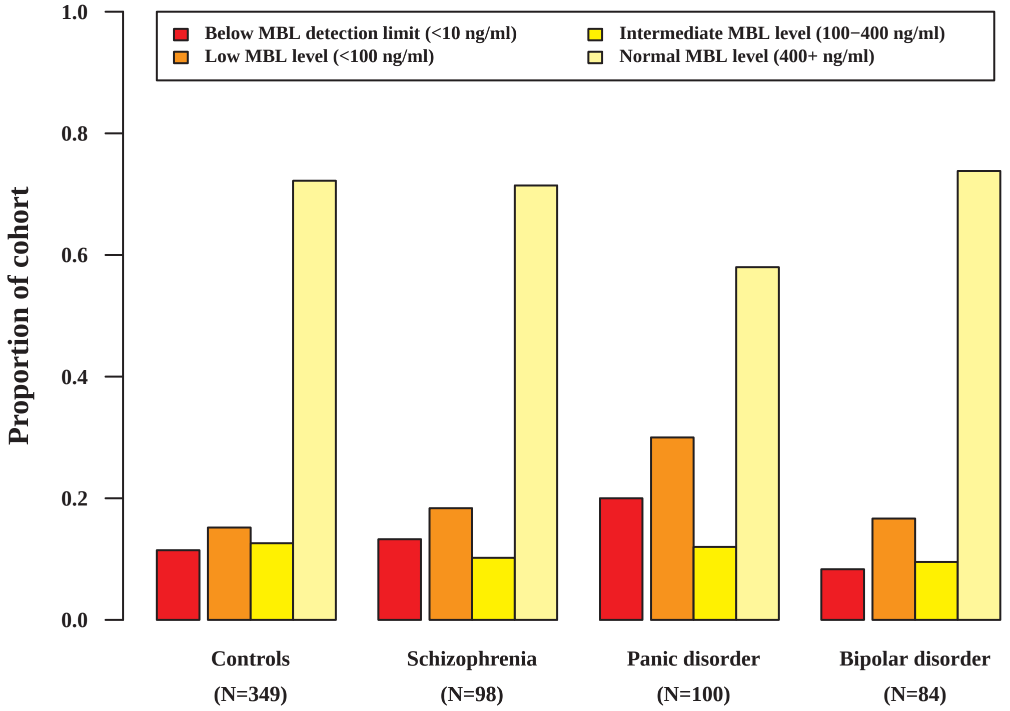


MBL serum concentration on *MBL2* two-marker genotypes. Box-and-whiskers plot of MBL serum concentration divided on two-marker genotypes (YA/XA/YO) for 349 controls, 98 patients with schizophrenia, 100 patients with panic disorder without a history of bipolar disorder, and 84 patients with bipolar disorder. Before logarithmic transformation, concentrations were measured in ng protein per ml serum. The black dotted lines indicate: below MBL detection limit (<10 ng/ml), low MBL level (<100 ng/ml), intermediate MBL level (100–400 ng/ml) and normal MBL level (>400 ng/ml). The gray dotted lines indicate the classification used in Olesen et al. (2006).

genotypes significantly increases model fit and thus should be preferred. In patients with bipolar disorder we found no difference in MBL level compared with controls, and this conclusion is independent of adjustment for the genetic variation of *MBL2*.

To enable comparison between the patient groups (referred to as contrasts) we ran two more models—one with and one without haplotype adjustment—with all samples included at the same time by using a four levelled phenotype variable. The results are shown in table 4.4. We see that without haplotype adjustment, patients with panic disorder have significantly lower MBL levels than each of the other three groups and that the other groups have equivalent levels. However, when adjusting for the effect attributed to *MBL2* multilocus genotypes—modeled as an additive genetic model in number of haplotypes—we observe that patients with schizophrenia have highly significantly increased MBL levels compared to all other three groups. The smallest contrast for schizophrenia were to bipolar disorder. In the adjusted model, as indicated above, the p-value for the contrast comparing patients with panic disorder to controls and to bipolar disorder was increased by an order of magnitude and stayed just slightly below the threshold of nominal significance. Patients with bipolar disorder continued not being significantly different from controls, though the coefficient increased vaguely by the adjustment for genetic markers.

Figure 4.6



Levels of MBL serum concentration. Bar plots of MBL serum levels for each of the four subject groups. The three yellowish bars sums to one within each cohort corresponding to the categories: low MBL level (<100 ng/ml), intermediate MBL level (100–400 ng/ml), and normal MBL level (>400 ng/ml). The separated red bars indicate the proportion of measures that was below the detection limit (<10 ng/ml) within each group. The proportion of measures below the detection limit was significantly higher (P=0.027) in patients with panic disorder as compared with controls, and the odds ratio was $_{1.1}\,1.9\,_{3.5}$. Furthermore, the odds of having a low MBL level (<100 ng/ml) was significantly increased (P=0.0008) for this group of patients with an odds ratio of $_{1.4}\,2.4\,_{4.0}$ compared to controls.

MASP-2 log-gaussian regressions

Along the same lines as the Tobit regression for MBL, we ran the log-gaussian linear regressions with inclusion of the four-levelled phenotype variable mentioned above and results are shown in table 4.5. Again, results within individual patient groups can be found in Table 4 of both paper 1 (Foldager et al., 2012) and paper 2 (Foldager et al., 2014). We ran a new forward inclusion procedure and ended up with more or less the combination of the models obtain from running each disorders separately—including the interaction between phenotype and the genetic variant of *MASP2*. Also the somewhat surprising effect on MASP-2 levels of the variants in *MBL2* exon 1 pertained to this combined analysis. The effect of these variants is a statistically significant increase of the MASP-2 level, i.e. the opposite direction of the effect these nonsynonymous variants have on the expression of MBL. It should be noted that the separate models for panic disorder and bipolar disorder (see Table 4 in paper 2) included the YA two-marker haplotype. However, as the O allele of *MBL2* exon 1 (i.e. any of B, D and C variants) is identifiable with the YO two-marker haplotype, and as the effect of the YA was a decrease of MASP-2, there is no contradiction in this. In the separate analysis for patients with schizophrenia, the final model also included counts of O alleles. To enable reasonably easy calculation of contrast tests

Table 4.3

Quantiles of MBL and MASP-2 serum concentration. Comparisons with the controls were carried out using an extended version of the usual median test (Conover, 1999). The p-values were based on Monte Carlo simulations (Hope, 1968) using 1e8 replicates. MBL concentrations below the detection limit were set equal to this 10 ng/ml limit.

		Quantile	Quantile (χ^2, p-value)		
	Probability	Controls	Schizophrenia	Panic disorder	Bipolar disorder
MBL	0.1	10	10 (0.24, 0.72)	10 (4.90, **0.031**)	10 (4.90, **0.031**)
	0.25	319	305 (0.014, 1)[b]	54 (11.2, **0.0011**)	309 (0.057, 0.89)
	0.5[a]	1133	1133 (6e-4, 1)[b]	704 (4.59, **0.041**)	1307 (0.26, 0.63)
	0.75	2460	2660 (0.83, 0.43)	1629 (3.31, 0.088)	2333 (0.071, 0.89)
	0.9	3809	4643 (0.66, 0.45)	2921 (1.30, 0.27)	4285 (0.98, 0.42)
MASP-2	0.1	260	229 (1.20, 0.35)	127 (39.3, **1.0e-8**)[c]	165 (32.3, **3.1e-7**)
	0.25	332	317 (0.83, 0.43)	210 (24.2, **1.9e-6**)	218 (48.4, **1.0e-8**)[c]
	0.5[a]	417	425 (0.58, 0.49)	331 (25.9, **4.1e-7**)	299 (25.8, **4.4e-7**)
	0.75	517	640 (4.65, **0.034**)	418 (13.0, **3.4e-4**)	403 (6.32, **0.016**)
	0.9	664	859 (9.55, **0.0031**)	549 (5.17, **0.023**)	596 (1.04, 0.33)

[a] The median.

[b] P=1 either means that a number close but less than 1 were rounded or that all of the replicates were at least as extreme as the observed. Matters nothing for the conclusion but we do not really believe that the p-value can be exactly one—at least not in this setting.

[c] P=1.0e-8 from 1e8 replicates means that none of the replicates were more extreme than the observed so in reality the p-value may well be smaller.

between patient groups, we also ran a model without the interaction term. Moreover, p-values from testing contrasts without adjustment for genetic effects were calculated. The p-values after adjustment for the genetic effects were 2–4 orders of magnitude lower than those obtained without this adjustment.

The serum concentrations of MASP-2 in patients with panic disorder as well as in patients with bipolar disorder were significantly lower than in controls, see table 4.5. This can also be seen in figure 4.3, figure 4.2 and from the quantiles in table 4.3. The reducing effect of the *MASP2* variant was a bit more pronounced in patients with schizophrenia and in patients with panic disorder than in controls—this is the interaction effect. The interaction was not significant (P=0.07) for patients with bipolar disorder but this may be a question of lower power, as serum concentration was only available from 84 of these subjects, contrary to the 98 and 100 patients with schizophrenia and panic disorder, respectively. The p-values from contrast tests show that the MASP-2 levels of patients with schizophrenia overall are about the same as in controls. The interaction model shows, however, that patients with schizophrenia that are carriers of the G allele of D120G in *MASP2* have significantly lower MASP-2 serum concentrations than controls whereas carriers of the wild type A allele have somewhat (and nominally significant) higher MASP-2 levels. So here the interaction was not merely a question of the effect size but also the direction of the effect. The size of the effect on the logarithmic scale can be seen in figure 4.2 and

Table 4.4

MBL Tobit regressions and contrasts between phenotypes. Association of MBL serum concentration with phenotype—a four levelled factor: schizophrenia, panic disorder, bipolar disorder and controls. Results from testing contrasts between phenotypes are given with and without adjustment for the additive effects of carrying 0, 1 or 2 copies of each specific *MBL2* haplotype.

Models			Contrast p-values		
	Parameters	$_{L95}\,Coef\,_{U95}$	Controls	Panic disorder	Bipolar d.
Phenotype only					
	Intercept[a]	$_{6.13}\,6.37\,_{6.60}$			
	Schizophrenia	$_{-0.51}\,-0.01\,_{0.48}$	0.96	**0.010**	0.79
	Panic disorder	$_{-1.32}\,-0.83\,_{-0.33}$	**0.0011**		
	Bipolar disorder	$_{-0.45}\,0.08\,_{0.60}$	0.78	**0.0061**	
Haplotype model[b]					
	Intercept[c]	$_{8.17}\,8.34\,_{8.52}$			
	Schizophrenia	$_{0.29}\,0.50\,_{0.71}$	**2.8e-6**	**5.1e-8**	**0.0048**
	Panic disorder	$_{-0.44}\,-0.23\,_{-0.02}$	**0.034**		
	Bipolar disorder	$_{-0.10}\,0.12\,_{0.33}$	0.30	**0.013**	
	LYPA	$_{-0.54}\,-0.31\,_{-0.08}$			
	LYQA	$_{-0.12}\,0.03\,_{0.17}$			
	LXPA	$_{-1.41}\,-1.27\,_{-1.13}$			
	LYPB	$_{-3.68}\,-3.51\,_{-3.34}$			
	LYQC	$_{-4.02}\,-3.58\,_{-3.13}$			
	HYPD	$_{-2.49}\,-2.28\,_{-2.07}$			

[a] Controls.

[b] The additive effect of carrying 0, 1 or 2 copies for each of the specific haplotypes.

[c] Controls with HYPA/HYPA multilocus genotype, i.e. 2 copies of the HYPA haplotype.

in the Supplementary Table 2 of paper 1 (subsection 6.1.1). In patients with panic disorder and bipolar disorder, the MASP-2 levels were lower than in controls and, of course (given the results just mentioned), also lower than in patients with schizophrenia. This conclusion is independent of which allele of D120G they carry but for G allele carrying patients with panic disorder, the difference was even more pronounced.

Table 4.5

MASP-2 linear regressions and contrasts between phenotypes. Association of MASP-2 serum concentration with phenotype as a four levelled factor. Results from testing contrasts between phenotypes are given with and without adjustment for *MASP2* and *MBL2* markers.

Models		Contrast p-values		
Parameters	$_{L95}\,Coef\,_{U95}$	Controls	Panic disorder	Bipolar d.
Phenotype only				
Intercept[a]	$_{5.97}\,6.02\,_{6.07}$			
Schizophrenia	$_{-0.03}\,0.07\,_{0.17}$	0.16	**7.2e-10**	**3.5e-8**
Panic disorder	$_{-0.43}\,-0.33\,_{-0.23}$	**2.4e-10**		
Bipolar disorder	$_{-0.41}\,-0.30\,_{-0.19}$	**4.8e-8**	0.691	
Adjusted model[b]				
Intercept[c]	$_{5.98}\,6.02\,_{6.07}$			
MASP2	$_{-0.77}\,-0.67\,_{-0.56}$			
Schizophrenia	$_{-0.02}\,0.07\,_{0.16}$	0.10	**2.1e-13**	**2.2e-10**
Panic disorder	$_{-0.44}\,-0.35\,_{-0.26}$	**3.8e-14**		
Bipolar disorder	$_{-0.40}\,-0.31\,_{-0.21}$	**3.6e-10**	0.47	
MBL2 O[d]	$_{0.08}\,0.13\,_{0.18}$			
		Wald test		
Interaction model[e]		statistics	p-values	
Intercept[c]	$_{5.96}\,6.01\,_{6.06}$			
MASP2	$_{-0.69}\,-0.56\,_{-0.42}$	-8.25	9.6e-16	
Schizophrenia (SZ)	$_{0.01}\,0.11\,_{0.20}$	2.21	0.027	
Panic disorder (PD)	$_{-0.41}\,-0.32\,_{-0.23}$	-6.83	2.1e-11	
Bipolar disorder	$_{-0.40}\,-0.31\,_{-0.21}$	-6.36	4.0e-10	
MBL2 O[d]	$_{0.08}\,0.13\,_{0.18}$	5.03	6.4e-7	
SZ : *MASP2*[f]	$_{-0.59}\,-0.30\,_{-0.02}$	-2.12	0.035	
PD : *MASP2*[f]	$_{-0.69}\,-0.37\,_{-0.05}$	-2.30	0.022	

[a] Controls.

[b] The result from testing the reduction from the interaction model to this simpler model with an F-test was: $F_{623,625} = 4.1$, P=0.017. Therefore, this simpler model will not give a fit which is too different from the fit using the model including the interaction terms.

[c] Controls with A/A of both the D120G marker in *MASP2* and the exon 1 marker in *MBL2*, i.e. 2 copies of the A allele.

[d] The additive effect of carrying 0, 1 or 2 copies of a nonsynonymous variant O in exon 1, i.e. any of the B, D and C variants. Note that the count of O alleles is identical to the count of YO two-marker haplotypes.

[e] The p-values here are from Wald tests (H_0: coefficient=0) evaluated in a t-distribution with degrees-of-freedom ($d.f.$) equal to the difference between the number of subjects and number of parameters in the model, i.e. $d.f. = 349 + 98 + 100 + 84 - 8 = 623$.

[f] Here ”V1 : V2” represents the interaction effect between V1 and V2.

4.2 Results from paper 3: G×E simulation study

We decided to simulate from all sixteen possible combinations of two minor allele frequencies ($MAF \in \{0.3, 0.4\}$) and two risk ratios ($RR \in \{1, 1.2\}$) between high risk and low risk homozygote genotypes for each DPL, two odds ratios ($OR \in \{2, 5\}$) related to a one unit increase of the DPE, and two exposure proportions given by a probability parameter ($\pi \in \{0.25, 0.5\}$) of a binomial environmental distribution ($bi(1, \pi)$ distribution, i.e. a Bernoulli distribution). We currently only consider models with two DPLs and one DPE and we use the GEM model (see subsection 3.2.3) and no epistatic changes of the penetrances (Pinelli et al., 2012). No extra noise from non-predisposing environmental factors were included and we use the same parameters for both DPLs with $W = 1$ (i.e. a dominant genetic model). Furthermore, we use a fixed sample size of 10,000 with equally many affected and unaffected individuals, and assumes the expected disease prevalence to be 1%. A flowchart of the simulation procedures is shown in Figure 4.7.

Figure 4.7

Flowchart for the G×E simulation study.

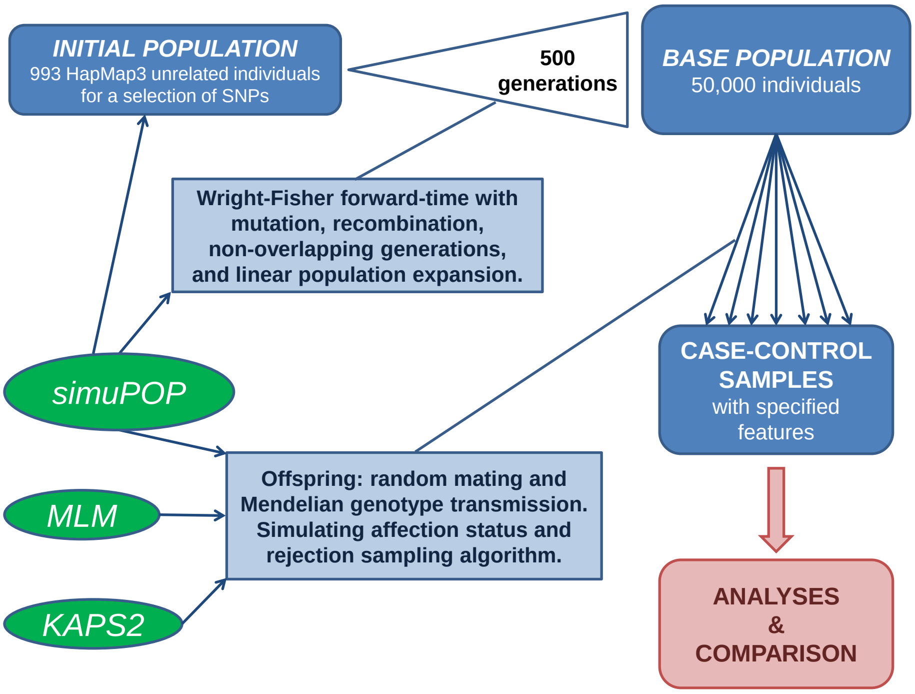

INITIAL POPULATION
993 HapMap3 unrelated individuals for a selection of SNPs
500 generations
BASE POPULATION
50,000 individuals
Wright-Fisher forward-time with mutation, recombination, non-overlapping generations, and linear population expansion.
simuPOP
CASE-CONTROL SAMPLES
with specified features
MLM
Offspring: random mating and Mendelian genotype transmission. Simulating affection status and rejection sampling algorithm.
KAPS2
ANALYSES & COMPARISON

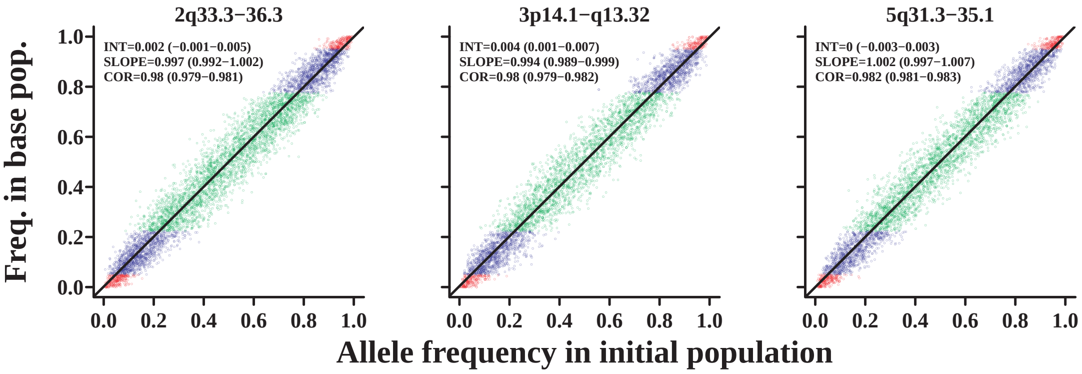
Table 4.6 shows which SNPs were chosen as DPLs, their allele frequencies in the initial population, and their allele and genotype frequencies in the base population. The site frequency spectrum plots of allele frequencies in the base population versus the initial population for the 3 regions containing DPLs are shown in 4.8. The plots for the three other chromosomal regions are shown in Figure 2 of paper 3. The reduction given by excluding SNPs according to $P(m/m) > 0.05$ is indicated by green coloured points. The extra SNPs added by the use of the less stringent criterion MAF>0.05 are the blue points whereas red points are those exclude by both criteria. There is no indication of serious problems.

Table 4.6 DPL allele and genotype frequency

			Genotype frequency[a]		
ID (chr)	MAF_{init}	MAF_{base}	M/M	M/m	m/m
rs4257797 (5)	0.34	0.30	0.49	0.42	0.090
rs1781740 (6)	0.23	0.30	0.49	0.42	0.089
rs2941399 (6)	0.38	0.40	0.36	0.48	0.16
rs7000415 (8)	0.41	0.40	0.36	0.48	0.16

[a] M=major allele, m=minor allele

Figure 4.8



Site frequency spectrum. Green points are SNPs for which the frequency $P(m/m)$ of the minor homozygote is above 0.05, blue points are the extra points obtained using a limit of MAF>0.05, and red points are those excluded by both limits. That is, all SNPs is the union of green, blue and red points. The slope line is the least squares fit using all points.

4.2.1 BOSS and BOOST results

Summary statistics of p-values obtain using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios were calculated for: 1) single-marker χ^2_2 genotype-based tests from BOOST (Figure 3 in paper 3); 2) single-marker additive tests adjusted for DPE main effect (Wald tests) from BOSS (Figure 4 in paper 3); 3) two-way interaction tests adjusted for SNP and DPE main effects (Wald tests) from BOSS (Figure 5 in paper 3). Furthermore results from two-way SNP-SNP interaction tests adjusted for main effects (χ^2_4) from BOOST are summarized in Figure 6 of paper 3.

The figures indicate that the simulated samples adhere to the simulated models with respect to genotypic main effects: RR=1.0 corresponding to no effect and RR=1.2 corresponding to a small main effect (same for both DPLs). There are no noticeable effects of varying the other parameters (MAF, OR and prevalence π of environmental exposure) except from the statistics concerning the G×E two-way interactions between SNPs and the DPE with adjustment for main effects (Figure 5 in paper 3). For this we note that the two-way interactions between DPLs and DPE are highly significant when the main effect of DPE is smaller (OR=2.0) and less prevalent ($\pi = 0.25$). Both increased prevalence of the environmental factor (DPE) and increase of its disease predisposing effect (OR=5.0) diminishes the significance of the interaction term.

It is noticeable that none of the minima (blue triangles and '+'s in Figure 3, 4 and 5 of paper 3) are above the Bonferroni adjusted threshold in the scenarios where RR=1.0 whereas this is the case for a small set of markers when RR=1.2. Interestingly those above the threshold are not just markers in close proximity to the DPLs. Probably this is due to longer ranging LD. Note also, that the main effects (RR=1.2) would remain undetected in many of the samples when using the Bonferroni threshold. Adjusting for the environmental effect diminishes these p-values even more (Figure 4 in paper 3). In accordance with the simulated models this reduction is larger for larger environmental effect, i.e. more pronounced when the samples were sampled with OR=5.0 than with OR=2.0.

Finally, bar plots summarising BOOST G×G χ^2_4 genotype-based tests (two-way interaction between SNPs) are shown in Figure 6 of paper 3. Only test statistics >30 (P<4.9e-6) were used and the tests are adjusted for main effects of the interacting SNPs. The bars are the number of samples (out of 100) where the SNP was present in at least one SNP-SNP interaction with a test statistic above the threshold of 30. In agreement with the models simulated (no epistasis), no systematic patterns are apparent and the DPLs are not more often part of G×G interactions than the other SNPs.

4.3 Results from paper 4: Suicide study

None of the genetic markers were significantly associated with suicidal behaviour in the basic single-marker tests. No other results of interest was found for the SNP rs13868494 in *TPH2* or for the X chromosomal marker *MAOA*uVNTR, see paper 4 (Buttenschøn et al., 2013). We will therefore here focus on results concerning the SNP rs1800532 in *TPH1* and the serotonin transporter locus in the 5' promoter region of *SLC6A4*.

4.3.1 Interactions in *TPH1*

For rs1800532 in *TPH1*, all three two-way interactions between gender, age-group and the additive term (A_i, see subsection 2.3.3) were statistically significant: gender by age-group (P=0.0022), gender by the additive term (P=0.014), and age-group by the additive term (P=0.00057). Thus, the effect of carrying the minor allele depended on both gender and age-group. The three-way interaction was not significant. For further investigation of these interactions, we calculated the OR (per minor allele) separately for each gender and age-group, see table 4.7 (corresponds almost to Table 3 in paper 4). In male subjects we observed a clearly significant protective effect of the minor allele in the youngest individuals whereas it tended to be a risk factor for the two other age-groups. In contrast to this, the effect in the oldest females was towards protection. From age-group specific conditional logistic regressions (gender stratified) we observed that the protective effect holds in general for subjects less than 35 years of age.

4.3.2 Results for *SLC6A4*

We will only consider the tri-allelic marker obtained by combining 5-HTTLPR and rs25531 (see subsection 2.2.3) and we will focus on results obtained after collapsing the S and L_G alleles, i.e. the functional activity genotype classes: SS+SL_G+L_GL_G (low expression), SL_A+L_GL_A (medium expression), and L_AL_A (high expression). The results for 5-HTTLPR (rs4795541) without rs25531 can be seen in Table 4 of paper 4.

Table 4.7

***TPH1* gender and age-group specific effects.** Results using the additive model for the SNP rs1800532 in *TPH1* within each combination of gender and age-group (<35, 35–49, ≥50 years). The last column contain results from gender stratified analyses within each age-group. Odds ratios are per copy of the minor allele.

P $_{L95}\,OR\,_{U95}$	Female	Male	Gender stratified
Age-group			
<35	0.10 $_{0.3}\,0.6\,_{1.1}$	**0.0064** $_{0.4}\,0.6\,_{0.9}$	**0.0011** $_{0.4}\,0.6\,_{0.8}$
35-–49	0.17 $_{0.7}\,1.1\,_{1.7}$	0.13 $_{0.9}\,1.6\,_{2.8}$	0.21 $_{0.9}\,1.2\,_{1.7}$
≥50	0.044 $_{0.5}\,0.7\,_{1.0}$	**0.038** $_{1.0}\,1.6\,_{2.5}$	0.86 $_{0.7}\,1.0\,_{1.3}$

Since the dominance effect was significant (P=0.0065), we used the genotypic model containing joint additive and dominance effects. There was a tendency of interaction between age-group and genotype (P=0.077) and a nominally significant interaction between age-group and the dominance effect (P=0.049). In table 4.8 (corresponding to Table 5 of paper 4) this tendency was explored by age-group separated comparisons between the medium expression class (SL_A+L_GL_A) and each of the low and high expression classes. From this we observe that the interaction with the dominance parameter is expressed as an over-dominance effect, with the medium expression heterozygote genotypes raising the risk more than the high expression homozygote genotype (L_AL_A). The age-specific calculations indicates that this effect may be more pronounced for the youngest individuals (<35 years).

Table 4.8

***SLC6A4* genotype-based results.** Results (odds ratios) from comparing the heterozygote genotype SL_A+L_GL_A (medium expression), with each of the two homozygote genotype classes in *SLC6A4*: SS+SL_G+L_GL_G (low expression) and L_AL_A (high expression). The effects were estimated using conditional logistic regression stratified on gender and age-group (All) or stratified on gender (separate age-groups).

P $_{L95}\,OR\,_{U95}$	SL_A+L_GL_A vs. SS+SL_G+L_GL_G	SL_A+L_GL_A vs. L_AL_A
Age-group		
All	**0.011** $_{1.1}\,1.7\,_{2.6}$	0.079 $_{1.0}\,1.4\,_{1.9}$
<35	**0.019** $_{1.2}\,2.8\,_{6.5}$	0.063 $_{1.0}\,1.9\,_{3.6}$
35-–49	0.16 $_{0.8}\,1.7\,_{3.7}$	**0.019** $_{1.2}\,2.4\,_{4.8}$
≥50	0.37 $_{0.7}\,1.4\,_{2.6}$	0.45 $_{0.5}\,0.8\,_{1.4}$

4.4 Results from paper 5: Slynar locus

Paper 5 exemplifies many of the aspects considered in section 3.1: genotype-based and allele-based single-marker analysis; testing for HWE; calculation of LD; different genetic models; haplotype analysis. Furthermore, methods for microsatellites (multi-allelic markers) are used in combination with methods for SNPs. Also, correction for multiple testing, the use of a replication sample, and a meta-analysis combining results from three studies are considered. In addition to the publication (Buttenschøn et al., 2010), these results were also presented by a talk at the XVIIth World Congress on Psychiatric Genetics (Foldager et al., 2009a).

We will not go into many details here, though, and given our conclusion regarding the use of allele-based tests (see subsection 3.1.3), we will concentrate on genotype- and haplotype-based results. Moreover, no convincing results were found for patients with schizophrenia and we will therefore only show results concerns patients with bipolar disorder. The genotype and allele counts for the genotyped SNPs are shown in the supplementary table include immediately after paper 5 (subsection 6.5.1) in this thesis and referred to as *Table 2 in the Supplementary material* in Buttenschøn et al. (2010) (in fact there is no supplementary table 1 so there is a typo here).

Concerning Hardy-Weinberg equilibrium we found that a few of the microsatellites deviated from Hardy-Weinberg proportions. The largest deviation was for marker m5 in controls, with a p-value of 0.002. The other deviations are barely worth mentioning but can be found in the paper. Some of the markers were in strong LD, see Figure 4 of paper 5. Most markedly for m9 and m10 with r^2s of 0.99 and 0.98 in controls but these markers were also physically very close. Haplotypes of microsatellite marker m8 and SNP marker m9, which showed the largest difference in frequencies between patients with bipolar disorder and controls (13% and 7%, respectively), were also in strong LD with each other ($r^2 = 0.95$).

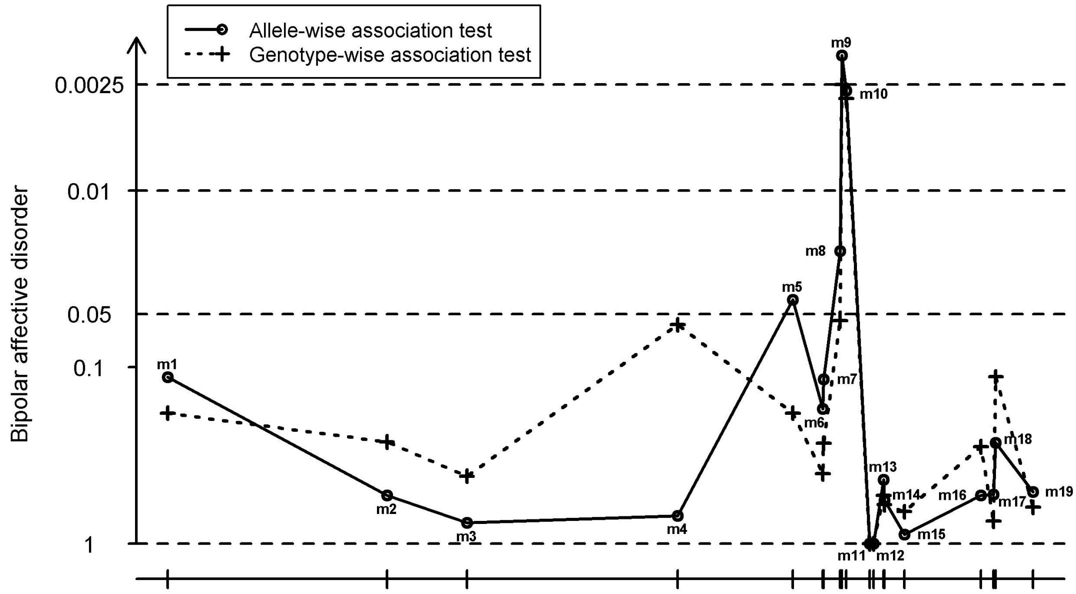
4.4.1 Single-marker analysis

Two markers, m9 and m10 were significantly associated with bipolar disorder with genotype-based p-values of 0.002 and 0.003, respectively. The minor allele (T in m9 and G in m10) was overrepresented among cases for both markers: 14% of the patients with bipolar disorder carried the T-allele at m9 compared with 7% of the controls, and 14% of the patients with bipolar disorder carried the G-allele at m10 compared with 8% of the controls.

P-values on a base-10 logarithmic scale from single-marker tests are shown in figure 4.9 corresponding to Figure 2 in paper 5 (Buttenschøn et al., 2010). The impact of m9 and m10 on disease risk was assessed by logistic regression. The saturated models (genotype-based association) were superior to the null model but did not fit significantly better than the corresponding additive and dominant models. A recessive genetic model was not supported. Based on Akaike's Information Criterion the additive genetic model was chosen and the OR per minor allele was found to be ${}_{1.3}\,2.0\,{}_{3.2}$ for m9 and ${}_{1.3}\,2.0\,{}_{3.1}$ for m10. The additive effect is multiplicative on the OR scale (exponentiated difference of log odds). Thus, the OR for homozygous carriers of the minor allele is the square of these odds ratios, i.e. 4.0 for both markers (m9 and m10).

Replication and refinement of association between bipolar disorder and the Slynar locus was the main purpose of the study. This corresponds to a main null hypothesis saying that none of the 11 markers within the Slynar region (m5-m15) are associated with bipolar disorder. The genotypic associations with m9 and m10 both survive correction for this family of tests

Figure 4.9



Single-marker association tests. Genotype-based and allele-wise association tests in the Danish bipolar disorder sample with p-values plotted on a base-10 logarithmic scale.

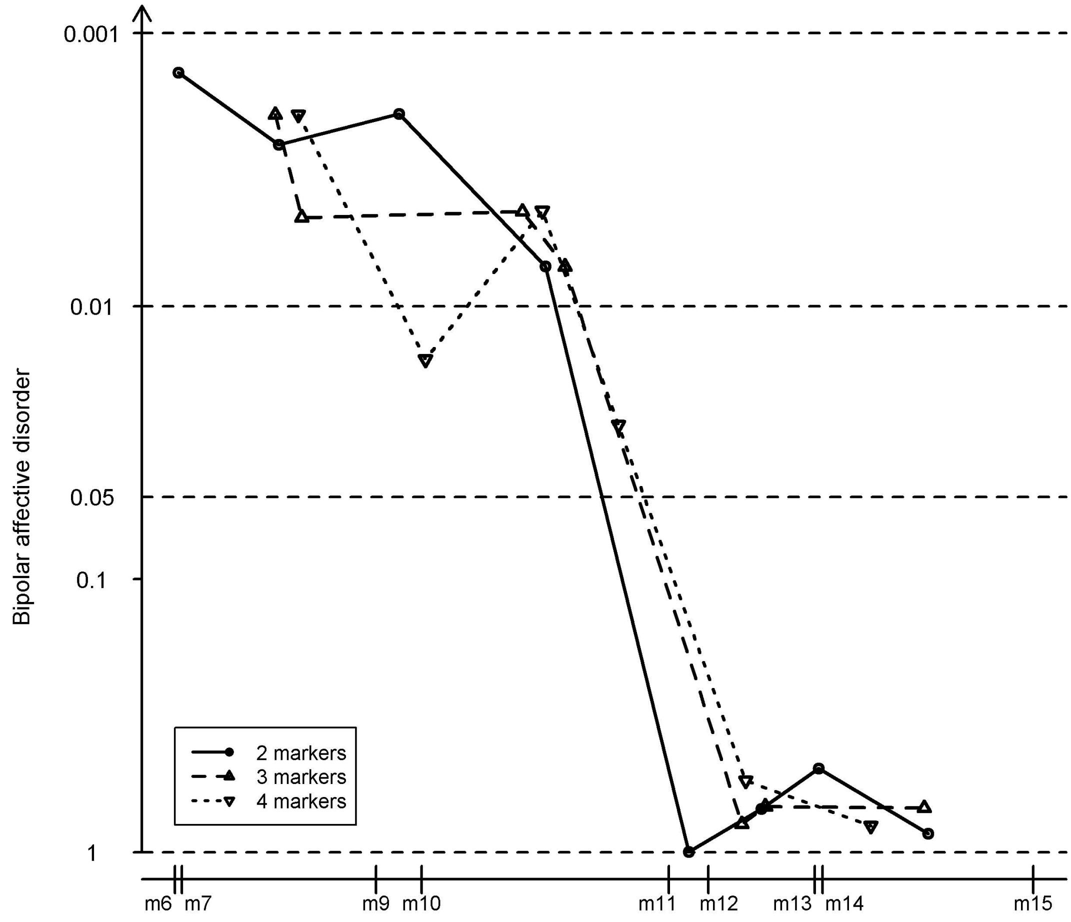
by Hommel's procedure (Hommel, 1988) and the corrected p-values were 0.025 and 0.030, respectively. If the family of tests is broadened to be all 22 allelic tests, then the corrected p-values are 0.052 and 0.063, and the results are just above the border of significance. Using instead the FDR method by Benjamini et al. (1995), these latter corrected p-values were instead both 0.033 and thus still indicating significant association of markers m9 and m10 with bipolar disorder.

Furthermore, m9 showed genotypic association with bipolar disorder in the Scottish sample (P=0.03) and in the combined Danish and Scottish sample (P=0.008). Similar to the Danish and the UK sample (Kalsi et al., 2006), the minor allele was overrepresented in cases compared with controls in the Scottish sample. In the combined Danish, Scottish and UK sample (918 patients with bipolar disorder and 946 controls in total), a meta-analysis of m9 using the additive genetic model showed an OR of $_{1.2}\,1.5\,_{1.9}$ per minor allele, which was clearly statistically significant (P=0.0003). Correspondingly, the OR for homozygous carriers of the minor allele was OR=2.2.

4.4.2 Haplotype analysis

The distribution of several two-, three- and four-marker haplotypes were significantly different (P<0.01) between patients with bipolar disorder and controls, see figure 4.10 (Figure 3 in paper 5). The most significantly associated two-marker haplotype included m6-m7 (P=0.001). This haplotype association was primarily caused by differences in the frequencies of two haplotypes (C-C and A-G). The C-C haplotype seems to be a risk haplotype ($P_{\mathrm{local}} = 0.0024$) with frequencies of 13.9% and 7.6% in patients with bipolar disorder and controls, respectively, whereas the A-G haplotype seems to be a protective haplotype ($P_{\mathrm{local}} = 0.0054$) with frequencies of 0 (unobserved) and 2.1%, respectively. The most significantly associated three- and four-marker haplotypes also included m6-m7. However, several of the remaining significantly associated haplotypes involved marker m9 and m10 and thus supported the results from the single-marker analysis.

Figure 4.10



Haplotype association tests. Two-, three- and four-marker haplotype association analysis of SNPs in the Danish sample of patients with bipolar disorder. The p-values are plotted on a base-10 logarithmic scale.

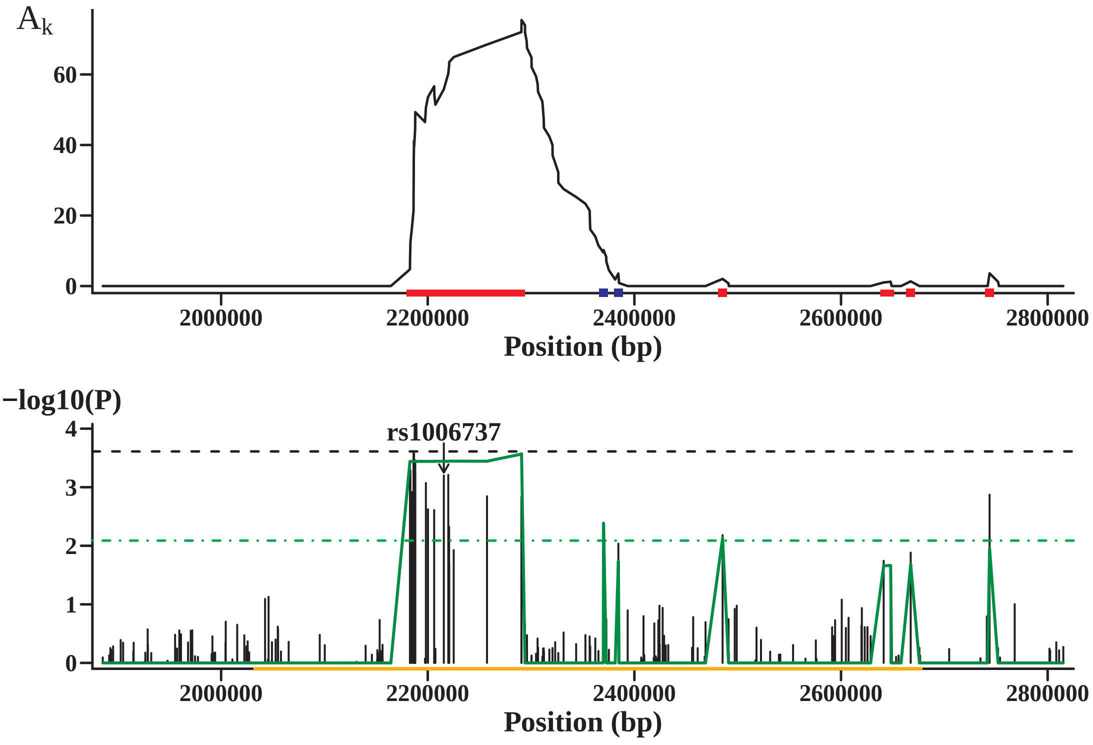
4.5 Results from paper 6: Landscape of *CACNA1C*

The results shown in figure 4.11 are from single-marker trend tests and from running the Landscape method for non-independent variables using Approach 2, see subsection 3.5.3 and Section 3.2 of paper 6. Here we defined $Z_k = \log(\frac{\alpha}{p})$ (see Example 3.2 in Section 3 of paper 6), where α=0.05 and p are p-values from the single-marker trend test.

We Bonferroni corrected the threshold of significance for multiple testing by dividing the significance level α with the mean number of maximal segments, as approximated by the average number of maximal segments in the permutation-samples. This average was $_{5.71}\,6.135\,_{6.56}$.

In conclusion, the Landscape method detects a clearly significant maximal segment for bipolar disorder in *CACNA1C* around rs1006737 spanning 108 kb and consisting of 26 SNPs.

Figure 4.11



Upper: Landscape plot against base pair (bp) position on chromosome 12p13.33 for $Z_k = \log(\frac{\alpha}{p_k})$ where $\alpha = 0.05$ and p_k are p-values from the single-marker test shown in the lower plot. Independent and dependent segments are indicated on the x-axis with red and blue bars, respectively. **Lower:** Results from single-marker trend tests carried out in *CACANA1C*, none of which are significant at level $\alpha = 0.05$ with Bonferroni correction for the 204 tests (threshold indicated with the black dashed line). The green line is from Landscape with Z_k as above and using the 999,999 permutation-based p-values. Bonferroni corrected threshold adjusted for the mean number of maximal segments (6.135) is indicated with the green dashed line. The orange line below the plot indicates the gene region of *CACNA1C*.

Chapter 5

Discussion, conclusions and perspectives

5.1 Discussion

5.1.1 Further investigation of the complement system

The complement system of activation is involved in the innate and adaptive immune defence and is activated through three pathways activated by different kinds of attacks (bacterial surfaces, antibody-antigen complexes and pathogen surfaces, respectively): the lectin, the classical and the alternative pathway. In the present thesis, we investigated the involvement of two components, MBL and MASP-2, which are both components from the lectin pathway. However, the complement system is much more complex and consists of more than 30 proteins which are either soluble in the blood or membrane-associated. The activation is followed by enzymatic reactions in a sequential cascade referred to as the complement activation pathways, see Figure 1 in Sarma et al. (2011) for an overview. In this process inactive zymogens (inactive enzyme precursors) are cleaved and activated.

Going through the full pathway would be too lengthy but to get a feel, we will briefly describe the lectin pathway below. Many of the components (complex proteins) are denoted by ”C” followed by a number and for some also a letter after the number. All three pathways converge at C3, the most abundant complex protein found in blood. Moreover, a number of regulating factors (e.g. inhibitors) for different parts of the system are known (Sarma et al., 2011; Mayilyan, 2012). Among these regulators are carboxypeptidases of which at least one encoding gene *CPXM2* (*carboxypeptidase X (M14 family), member 2*) has shown some evidence of association with cognitive decline in schizophrenia (Hashimoto et al., 2013). Havik et al. (2011) investigated regulators of the classical pathway and considered furthermore a set of brain-expressed regulators of complement activity (RCA). They found association between schizophrenia and the complement-control related genes *CSMD1* and *CSMD2* (*CUB and Sushi multiple domains 1 and 2*). Markers within *CSMD1* were also found to be genome-wide significantly associated with schizophrenia in the first PGC meta-analysis (Ripke et al., 2011) and have been identified as a target for regulation by miR-137 (Kwon et al., 2013). The neuropsychological effects of *CSMD1* were furthermore investigated, and it appears that it may be involved in mechanisms related to memory and learning (Donohoe et al., 2013).

The lectin pathway is initiated by ficolin-MASP-2 or MBL-MASP-2 protein complexes (interacting proteins). The MBL or ficolin lectin binds to pathogens and induces an auto-activation of MASP-2 which cleaves C4 into C4a and C4b. C4b attaches to the pathogens, which leads to binding of C2. Then MASP-2 cleaves C2 into C2a and C2b, and C2a attaches to C4b to form the C3 convertase C4bC2a. MBL and ficolin also complex with MASP-1, MASP-3 and with a truncated MASP-2 referred to as sMAP (Matsushita, 2010) or MAp19 (Sørensen et al., 2005), but the role of these three proteins in the lectin pathway is more uncertain. The (human) genes encoding the lectin proteins are: *MBL2* (encoding MBL), *FCN1*, *FCN2*, *FCN3* (encoding ficolin-1, -2 and -3 also known as M-, L- and H-ficolin, respectively), see Garred et al. (2009). MASP-1, MASP-3 and MAp44 are all three encoded by *MASP1* whereas both MASP-2 and MAp19 are encoded by *MASP2* (Degn et al., 2010).

We may therefore consider doing a pathway-based analysis using markers within the regions identified by genes involved in the activation cascade. This also exemplifies one way to limit the number of markers considered, thereby enabling a more thorough analysis of more complex interaction patterns.

5.1.2 Simulation of data

Many methods and software have been proposed for simulation of data from multiple disease SNPs. The best choice depends on the disease model one wants to simulate, how large regions should be simulated, sample sizes, number of SNPs etc. As already noted, Hoban et al. (2012) reviewed a larger collection (42 simulation packages) but their list is by no means exhaustive. Another source, which includes some possibilities for comparing pros and cons of the methods, is the web page of genetic simulation resources[29](Peng et al., 2013).

An interesting alternative not mentioned by Hoban et al. (2012) is the HAPGEN2 by Su et al. (2011) which uses data from reference panels of haplotype data like HapMap2, HapMap3 and 1000Genomes to obtain LD patterns similar to those observed in real data. It simulates multiple disease SNPs on a single chromosome and may also include G×G interactions by use of an R package *SimulatePhenotypes* (available only from the HAPGEN2 web page!) to simulate phenotypes for a set of genotype data. Nevertheless, it can not be used for simulation of environmental impacts on the disease risk and thus neither for G×E interactions.

A practical issue in connection with machine learning methods is the need of complete data, i.e. no missing genotypes or other measures. This is not needed when using logistic regression or other generalised linear models. The MB-MDR and logicFS softwares used for the simulation study in paper 3 are not exceptions from this rule.

Finally, optimal methods should take genotype probabilities and thereby allow for imprecision (or variation) of genotyping (and/or imputation) as well as avoiding the need for complete data. This may be worth having in mind when choosing further G×E methods for comparison.

5.1.3 The study on suicidal behaviour.

The involvement of *TPH1* (rs1800532) in suicidal behaviour has often been investigated but with inconclusive and contrary results (see references in paper 4). In accordance with many of the studies, we did not find a significant association between rs1800532 and completed suicide, but

[29] `http://popmodels.cancercontrol.cancer.gov/gsr`

the inclusion of interactions revealed a more complex picture which might be a reason to the contradicting results between studies. We found that the effect of carrying the A-allele depends both on gender and age-group. A clearly significant protective effect of the minor A-allele was observed in male subjects younger than 35 years, while the A-allele tended to be a risk factor for older male subjects. In contrast to this, a protective effect was observed for females in the oldest age-group.

Results from studies of the 5-HTTLPR genetic marker located within the SLC6A4 gene have also been conflicting. Exploratory interaction analyses in paper 4 (Buttenschøn et al., 2013) showed that the effects may depend on age-group. An elevated suicide risk was observed for heterozygous individuals between 35 and 49 years compared to homozygous individuals. Further exploration of this revealed that this effect was more pronounced in males than in females (results not shown). The interaction analyses of the tri-allelic marker in the serotonin transporter showed a statistically significant protective effect of the low expression genotypes for individuals below 35 years, and a statistically significant protective effect of the high expression genotype for individuals between 35 and 49 years. Interpreting these results is not obvious, and they may reflect some underlying unobserved factors. As far as we know, similar analyses have not been performed by others.

5.1.4 The slynar locus.

The Slynar locus (m5-m15, see Figure 1 in paper 5) was investigated by inclusion of two microsatellites and nine SNPs, six of which were tag SNPs (m6, m7, m9, m11, m12, m13), i.e. SNPs covering a larger region due to high LD. The results supported the presence of a susceptibility locus for bipolar disorder on chromosome 12q24.3 and specifically implicated a 50 kb region within the Slynar locus. The function of Slynar still appears unknown, however, and further functional studies are needed to clarify the function of the gene and the importance of this gene in bipolar disorder.

The two most significantly associated markers (m9 and m10) were also associated with bipolar disorder in a UK cohort. A very high linkage disequilibrium was observed between these two markers ($r^2 > 0.98$). The power to detect an odds ratio of 2 for carriers of one minor allele in an additive model was found to be 77% under assumptions of a disease prevalence of 1%, and a 14% frequency of the risk allele as observed for m9. The association of m9 was further confirmed by replication in a Scottish sample.

No markers in the WTCCC1 bipolar disorder GWAS (Wellcome Trust Case Control Consortium, 2007) or in the study by Sklar et al. (2008) showed any significant associations within 12q24. One of the markers (rs1706509) from the chip used in these studies is located in relative proximity of m9 (rs7133178). Despite the relatively short distance between these two SNPs (3208 bp), they are in very low LD, $r^2 = 0.005$. This might be one explanation why no significant association with rs1706509 appeared in these studies. Another explanation could simply be low power due to effect or sample size limitations.

5.1.5 Limitations

The main limitations are: Sample size, sample size, sample size, ...

Single-marker genetic effects in complex mental disorders are likely to be relatively small. Most of the samples considered in this dissertation are relatively small, and low power to detect real associations may therefore be the most severe limitation. Regarding doing interaction analyses this limitation is even more pronounced.

The use of a limited number of genetic markers on the hand means that the problem regarding multiple comparison and associated lower threshold of significance is less pronounced. So though using a small set of markers may be seen as a limitation of the studies, it may also have its pros.

In paper 1 and 2, large variations of serum concentrations may be another power issue but differences may also be much more pronounced from such quantitative traits. Furthermore, exclusion of other components from the complement pathway was a limitation of this study.

In the study on completed suicide (paper 4), amplification of DNA extracted from paraffin blocks gave some technical problems for the longer fragments and thus especially for 5-HTTLPR. The use of frozen tissue was less problematic, unless the tissue was very badly degraded before freezing. As a result, more samples were excluded not least in the analysis of markers from *SLC6A4*. Overall, 10 % of the suicide cases were excluded completely from the study due to degraded tissue.

For some of the control samples, there was a lack of information concerning demographics, environmental exposures and information of ancestry. This, of course, limits the possibilities of controlling for confounding or effect-modifying factors. This limitation can be hard to avoid partly due to economic restraints but also simply because it may be difficult to either recruit new controls or get the relevant measures or information for earlier collected control samples. The use of the Danish Newborn Screening Biobank (Norgaard-Pedersen et al., 2007) is to some extent an exception and one of the reasons why this is an exceptionally important source for Danish genetic research. Another limitation is the use of unscreened controls, like the medical students used in paper 4, and the use of standard controls with a possible different gender distribution than cases.

The dichotomised categorisation as patients with a specific disorder relies on observations of symptoms. Obviously this may be subject to subjectivity and differences as to how these symptoms are perceive by the individual. As an example the so-called positive symptoms in schizophrenia are audio-visual misconceptions (hallucinations) and thus can not be measured or checked. Furthermore symptoms overlap between disorders, and distinctions may not be as clear as the binary affected/not-affected, schizophrenia/bipolar etc. categorisation propose. Subcategories of the disorders do exist and might be used to introduce a finer partitioning of the outcome but might be with poor quality because many subjects might be given the general diagnosis rather than the more correct subcategory. Along the same lines the lack of information on course and severity of the disorders can be a limitation.

5.2 Conclusions

In the following subsection we will give the main conclusions from the six papers. Overall, we conclude that the awareness and possible inclusion of interactions may reveal relations that might otherwise have been overlooked. The same conclusion goes for the use of multi-locus methods and of methods summarising signals.

5.2.1 MBL and MASP-2

A very clear effect of higher MBL serum concentration in patients with schizophrenia was seen when adjusting for the variation in MBL ascribed to *MBL2* variants, see table 4.4. The level of MBL was also higher in patients with bipolar disorder but significantly lower than in patients with schizophrenia. At the same time, however, the lower quantiles were lower in patients with schizophrenia than in patients with bipolar disorder, see table 4.3, and specifically the median of MBL was exactly the same in patients with schizophrenia as in controls but almost 200 ng/ml higher in patients with bipolar. This is in accordance with the observation that the proportion of patients with schizophrenia in the MBL low-producing multilocus genotype groups (19%) was higher too, and with the fact that equally many patients with bipolar disorder (20%) were in the low group but fewer were in the intermediate group: 23% in contrast to 28% in both controls and patients with schizophrenia, see table 4.2.

Patients with panic disorder, on the other hand, had a remarkably and statistically significant lower serum concentration of MBL than the three other groups. Though MBL deficiency does not necessarily lead to development of clinical deficit symptoms, it is interesting that among patients with panic disorder, 30% had MBL deficiency according to the $<$100 ng/ml limit. This proportion was 15–18% in the three other groups, and the observation is in agreement with the higher frequency of panic disorder patients carrying *MBL2* diplotypes XA/YO and YO/YO that are known to be associated with low MBL levels (Garred et al., 2006; Heitzeneder et al., 2012).

MASP-2 serum concentration was lower in patients with schizophrenia than in controls for subjects carrying the D120G mutation but higher for wild-type carriers. This interaction was significant, see table 4.5. Patients with bipolar disorder and patients with panic disorder had equivalent MASP-2 levels, which were highly significantly lower than the concentrations seen in patients with schizophrenia and in controls. Carrying the D120G mutation further lowered the level in patients with panic disorder but not significantly more in patients with bipolar disorder. Interestingly, MASP-2 levels also depended significantly on variants in *MBL2* exon 1.

The differences in MBL and MASP-2 serum concentrations between controls and patients suffering from schizophrenia, bipolar disorder or panic disorder are intriguing, but the genetic analyses gave no definite answer as to why these levels differ. This may indicate that more insight into the aetiologies of mental disorders can be found by analysing the complement pathway of activation in greater detail. Since MBL deficiency is highly heterogeneous and associated with both infectious and autoimmune states, more research is needed to identify how the complement system could be associated with the mental disorders. Since the lectin pathway is very complex, a functional assessment may be relevant in addition to measurement of MBL and MASP-2. The absence of association with the functional variants in exon 1 may well be a power issue rather than lack of true association.

In conclusion, this study supports previous studies showing increased complement activity in patients with schizophrenia but indicates furthermore that changes in complement activity may be associated with other mental disorders as well. However, the direction depends on the diagnosis and may suggest aetiological heterogeneity among patients, underlining that multilocus genotypes have to be considered. It is apparent that inclusion of additional components from the complement system will be vital to further investigation of the association between psychiatric disorders and the activation pathway.

5.2.2 G×E simulation study

We were able to generate genotypic data with inclusion of an environmental factor impacting the penetrance via G×E interactions. To be totally confident, though, we still need to try to vary some of the other variables and not least try to include epistatic changes. Moreover, we need to further investigate how well the simulated samples comply with the penetrance model used for the generation of affection status and not least if the effects can be found by the more involved data mining and machine learning methods.

Not much more can be said at this early stage of the study, but it is fair to say that patience is a virtue when performing simulation studies. All sorts of unwanted obstacles should be expected: software not running, bugs in scripts (both your own and others), numerical problems, server breakdowns and so forth. Another problem is scalability of the methods with limits on the number of factors/covariates that can be included. These limits may be software specific in terms of restrictions defined in the programs or hardware induced by memory limits or processor capacities. On the other hand, one of the advantages of using machine learning methods is the possibility to search for higher order interactions without being compromised by the need to adjust for multiple testing adjustment to a degree that the effect sizes or sample size have to be unrealistically large.

5.2.3 Suicide study

The study presented in paper 4 (Buttenschøn et al., 2013) is one of the largest studies on completed suicide. We investigated for association with five genetic markers located within four genes involved in the serotonergic system. Our findings suggest that none of these genetic variants are strong risk factors. Interaction analyses, however, indicated the importance of age and gender, see subsection 5.1.3. To reveal a better understanding of the genes involved in suicide, we suggest that future studies should include both genetic and non-genetic factors.

In agreement with the literature, the suicides from our study included more males than females. More females than males had a history of contact with a psychiatric hospital, however. In total 57 % of all suicide cases had a history of contact, and in most cases suicide was committed within one year since last contact. This is in agreement with the large population based study by Qin (2011).

5.2.4 Slynar study

Replication and refinement of association between bipolar disorder and the Slynar locus on chromosome 12q24.3 was the main purpose of the study in paper 5 (Buttenschøn et al., 2010). Two markers, m9 and m10 were significantly associated with bipolar disorder, and the most significantly associated marker, m9, was also associated with bipolar disorder in a UK cohort and in a Scottish replication sample. In a meta-analysis of these three cohorts, we found an odds ratio of 1.5 for carriers of one minor allele and OR=2.2 for carriers of two minor alleles.

The distribution of several two-, three- and four-marker haplotypes was also significantly different between patients with bipolar disorder and controls, see figure 4.10. The most significantly associated haplotypes included m6-m7. However, several of the remaining significantly associated haplotypes involved marker m9 and m10 and thus supported the results from the single-marker analysis.

In conclusion the exact replication of markers associated with disease status supports 12q24.3 as a region of functional importance in the pathogenesis of bipolar disorder. Since no SNPs analysed in the GWAS mentioned in subsection 5.1.4 were good proxies for the most significantly associated marker, the results in paper 5 also confirm the importance of focused genotyping.

5.2.5 Landscape method

We have developed a method to aggregate sequentially ordered statistics that may be applicable as a complementary method when searching for candidate regions, e.g. in whole-genome studies. We showed its potential by obtaining a statistically significant aggregated score for a region on *CACNA1C* using WTCCC bipolar data (Wellcome Trust Case Control Consortium, 2007) where the individual p-values were above even a region-based Bonferroni corrected threshold of significance—not to mention the genome-wide threshold of 5e-8 often used. Thus, the area might potentially have been suggested earlier by use of the Landscape method.

The assessment of significance in terms of p-values for scores of maximal segments was generally carried out using bootstrapping, and we showed examples of how to use these bootstrapped samples to correct for multiple testing. Compared to the correction needed when doing e.g. all single-marker tests in a region, the size of the adjustment needed in the Landscape method may be orders of magnitude lower. This was exemplified by the WTCCC example where a Bonferroni correction for 204 tests diminished to correction for a mean number of maximal segments estimated from the bootstrap samples to be just 6.135.

5.3 Perspectives

Identification of risk genes does potentially have an important impact on treatment in the future as they may point at drug targets and maybe pave the way for designing drugs with better treatment response and fewer adverse effects. Considerations of protein-protein interactions and G×G interactions seem unavoidable in this endeavour. The environmental background, including G×E interactions, may also play an important role for disease susceptibility and should be considered both for drug designing purposes but also with prevention in mind. The long-term perspective may be the ability to point at the most effective drug in advance of initiating treatment and thereby avoiding long and unpleasant series of trials of treatment with various non-effective drugs. An aspect of more thorough ethical concern which cannot be ignored is of course prenatal diagnosis and other interpretations of genetic risk factors in clinical practice.

An important area which we have not touched upon is the use of matched cases and controls. We did not use matched data in the studies presented in this thesis but we have had many considerations on extending the *logicFS* logic regression method to the case of a matched design. Matching is often used in epidemiological studies to ensure that cases and controls share certain characteristics, e.g. gender and age. This is probably less common in genetic studies but at least in Denmark the inclusion of information from registers has lead researchers to match cases and controls in such studies, see e.g. Borglum et al. (2013). It is not obvious, though, that this matching implies correlations between matched cases and controls with respect to genetics. Investigations of gene-environment interactions are probably more prone to bias from confounders but even in such cases it may sometimes be debatable whether matching leads to improvements (Faresjö et al., 2010). However, from a statistical point of view if a matched design was used then certainly matching should be taken into account when analysing the data. The need to stratify may also stem from other factors than the design, e.g. geographic location, hospital or laboratory differences in multi-site studies, batch effects and population stratification. Often conditional logistic regressions or stratified proportional hazards models are applied to handle this correlation. The *BOSS* method (Voorman et al., 2012) that we used for some calculations in paper 3 is actually an example of efficient computations in a setup allowing for correlated errors.

Within the framework of the original logic regression, it is fairly simple to handle 1:m matching (but not n:m). Actually, the documentation of the LogicReg software gives this as an example of how to implement a new scoring function by writing some lines of Fortran code and re-compiling the LogicReg package. Of course, the latter may be a technical hurdle but it can be done, and the implementation obtained by doing this fits the models by use of a stratified proportional hazard model.

Family-based data is a special kind of a matched design, and a version of logic regression adapted for the analysis of case-parent trio data was introduced by Li et al. (2009) and Li et al. (2010a). This method has been implemented in the R Bioconductor[27] package trio and uses a case-pseudo-control approach in which each case is matched to three pseudo-controls. The feature selection version *trioFS* later developed by Schwender et al. (2011a) has been added to trio and includes importance measures (VIMs).

We believe that it will be possible and relevant to also develop a feature selection (bagging) version of logic regression for usually matched designs—preferable for n:m matching but at least 1:m. Maybe the trick is to draw the bootstrap samples with replacement from the strata. We were considering simply building on top of the logicFS package and maybe borrow ideas from trio (including *trioFS*) but there are caveats if we consider using it on a larger scale (i.e. with many genetic markers) due to restrictions in the Fortran part of the software. There may therefore be more perspective in starting from scratch, so to speak. Doing so, efforts should be made to implement effective algorithms that utilise HPC, i.e. cluster computing. We have suggested to call such a version *conditional logic regression* (Foldager et al., 2010).

Chapter 6

Manuscripts

The thesis is based on the following manuscripts which are reproduced below:

1. **Foldager L**, Steffensen R, Thiel S, Als TD, Nielsen HJ, Nordentoft M, Mortensen PB, Mors O, Jensenius JC. *MBL and MASP-2 concentrations in serum and MBL2 promoter polymorphisms are associated to schizophrenia.* Acta Neuropsychiatrica 2012; **24**(4): 199–207.

2. **Foldager L**, Köhler O, Steffensen R, Thiel S, Kristensen AS, Jensenius JC, Mors O. *Bipolar and panic disorders may be associated with hereditary defects in the innate immune system.* Journal of Affective Disorders 2014; **164**: 148–154 (in progress), e-pub ahead of print 8 May 2014.

3. **Foldager L**, Als TD, Grove J. *Comparison of methods for genome-wide gene-environment interaction analysis.* Manuscript in preparation.

4. Buttenschøn HN*, Flint TJ*, **Foldager L**, Qin P, Christoffersen S, Hansen NF, Kristensen IB, Mortensen PB, Børglum AD, Mors O. *An association study of suicide and candidate genes in the serotonergic system.* Journal of Affective Disorders, 2013; **148**(2–3): 291–298.

5. Buttenchøn HN*, **Foldager L***, Flint TJ, Olsen IML, Deleuran T, Nyegaard M, Hansen MM, Kallunki P, Christensen KV, Blackwood D, Muir W, Straarup SE, Als TD, Nordentoft M, Børglum AD, Mors O. *Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3.* Psychiatric Genetics 2010; **20**(3): 93–101.

6. Wiuf C*, Pallesen JSM*, **Foldager L**, Grove J. *Landscape: A simple method to aggregate p-values and other stochastic variables without a priory grouping.* Manuscript in preparation.

*) Contributed equally to this study

6.1 Paper 1[30]

Acta Neuropsychiatrica
An International Journal of Translational Neuropsychiatry

Acta Neuropsychiatrica 2012: 24: 199–207
All rights reserved
DOI: 10.1111/j.1601-5215.2011.00618.x

MBL and MASP-2 concentrations in serum and *MBL2* promoter polymorphisms are associated to schizophrenia

Foldager L, Steffensen R, Thiel S, Als TD, Nielsen HJ, Nordentoft M, Mortensen PB, Mors O, Jensenius JC. MBL and MASP-2 concentrations in serum and *MBL2* promoter polymorphisms are associated to schizophrenia.

Objective: Causative relations between infections and psychosis, especially schizophrenia, have been speculated for more than a century, suggesting a hypothesis of association between schizophrenia and hereditary immune defects. Mannan-binding lectin (MBL) is a pattern-recognition molecule of the innate immune defence. MBL deficiency is the most common hereditary defect in the immune system and may predispose to infection and autoimmunity. Mannan-binding lectin serine protease-2 (MASP-2) is an MBL-associated serine protease mediating complement activation upon binding of MBL/MASP to microorganisms. The objective was to investigate if schizophrenia is associated with serum concentrations of MBL and MASP-2 or with genetic variants of the genes *MBL2* and *MASP2* encoding these proteins.
Methods: The sample consisted of 100 patients with schizophrenia and 350 controls. Concentrations of MBL and MASP-2 in serum were measured and seven single nucleotide polymorphisms known to influence these concentrations were genotyped.
Results: Significant association of disease with genetic markers was found in *MBL2* but not in *MASP2*. Significant difference in MBL serum concentration was found between patients and controls when adjusting for *MBL2* haplotypes. For concentrations of MASP-2, a significant interaction effect between a *MASP2* variant and disease was found. Interestingly, MASP-2 levels also depended significantly on variants in *MBL2* exon 1.
Conclusion: This study supports previous studies showing increased complement activity in patients with schizophrenia, indicates aetiological heterogeneity among patients and underlines that multilocus genotypes have to be considered when investigating effects on MBL level. It appears that inclusion of additional components from the system of complement activation is warranted.

Leslie Foldager[1,2], Rudi Steffensen[3], Steffen Thiel[4], Thomas Damm Als[1,5], Hans Jørgen Nielsen[6], Merete Nordentoft[7], Preben Bo Mortensen[8], Ole Mors[1], Jens Christian Jensenius[4]

[1]Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Denmark; [2]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark; [3]Department of Clinical Immunology, Aalborg University Hospital, Aalborg, Denmark; [4]Institute of Medical Microbiology and Immunology, Aarhus University, Aarhus, Denmark; [5]National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark; [6]Department of Surgical Gastroenterology 435, Hvidovre University Hospital, Hvidovre, Denmark; [7]Psychiatric Centre Copenhagen, University of Copenhagen, Copenhagen, Denmark; and [8]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark

Keywords: genetics; lectin pathway; mannan-binding lectin deficiency; mannan-binding lectin serine protease type 2; schizophrenia

Leslie Foldager, Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Skovagervej 2, DK8240 Risskov, Denmark.
Tel: +45 7847 1119;
Fax: +45 7847 1108;
E-mail: Leslie.Foldager@ps.rm.dk

Accepted for publication August 23, 2011

Significant outcomes

- The study support previous findings of increased complement activity in patients with schizophrenia.
- There was indication of aetiological heterogeneity among the patients.
- The results emphasise that multilocus genotypes should be used when examining for genotypic effects on MBL serum concentration.

[30] Acta Neuropsychiatrica 2012; **24**(4): 199–207. © Scandinavian College of Neuropsychopharmacology (SCNP). Reused with permission.

Limitations

Single-gene genetic effects in complex mental disorders are likely to be relatively small. Low power to detect real associations may therefore be the most severe limitation. The large variations of serum concentrations may be another power issue but differences may also be much more pronounced from such quantitative traits. Finally, the lack of ancestry restrictions to the sample of controls is considered a minor limitation. Information regarding demographics, on the other hand, would have been ideal in order to include possible environmental confounders or effect-modifiers.

Introduction

Causative relations between infections and psychosis, especially schizophrenia, have been speculated for more than a century (1). Schizophrenia has been associated with a number of autoimmune diseases and a 45% increase in risk for schizophrenia has been found for subjects with a history of autoimmune disease (2). Moreover, maternal infections during the embryonic stage or infections in early childhood are possible risk factors for psychosis (3–5). In a recent paper Håvik et al. (6) observe an association with schizophrenia for single nucleotide polymorphisms (SNPs) in genes encoding *CSMD1* and *CSMD2*. These genes encode proteins with a domain structure which is seen in some control proteins of the complement cascade, but also in a number of other proteins outside this system. Although not explored extensively it could be that the encoded proteins are influencing the activity of the complement system, e.g. a soluble form of the corresponding rat protein was tested positive for such activity (7). Furthermore they found associations with genes from the major histocompatibility complex (MHC) region on chromosome 6. This study and other recent results from large and combined studies showing association with genetic markers in the MHC region (8–10) are also consistent with a possible (auto)immune system connection. Hence a hypothesis of an association between schizophrenia and hereditary immune defects is suggested.

Mannan-binding lectin (MBL) is a pattern-recognition molecule of the innate immune defence. MBL deficiency is, with a prevalence of 10%, the most common hereditary defect in the human immune system and may predispose to infection and autoimmunity (11). As reviewed by Mayilyan et al. (12) studies have shown increased activity of the lectin pathway of complement activation in patients with schizophrenia, mainly in complexes with mannan-binding lectin serine proteases (MASPs). Two key components in this activation process are MBL and MASP-2 with the latter as main initiator of the lectin complement pathway (13). The genes encoding these proteins are *MBL2* located at 10q21.1 and *MASP2* located at 1p36.22 (UCSC Genome Browser hg18, March 2006, http://genome.ucsc.edu).

The molecular basis for MBL deficiency is reviewed in Garred et al. (14). Substantially decreased level of MBL is known to be associated with the presence of three non-synonymous mutations in exon 1 of *MBL2* while three polymorphisms from the promoter region explain much of the remaining variation in the serum concentration of MBL. Seven haplotypes formed by these six variants are common and correlate with different levels of MBL. Differences in haplotype frequencies may explain some of the variation in serum concentration seen between humans of different ancestral origin (14).

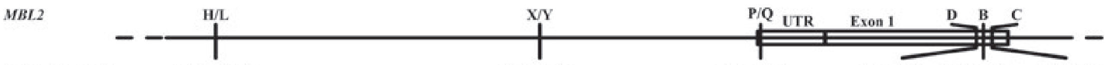
Aim of the study

The main objective of this study was to investigate in a Danish case–control sample if schizophrenia is associated with concentrations of MBL and MASP-2 in serum or with genetic variants of *MBL2* and *MASP2*. Subsequently we explored for a possible disease association with the protein levels after adjustment for the known effect of the polymorphisms.

Material and methods

Samples

From previous genetic studies a sample of 100 patients with schizophrenia was obtained. The patients were diagnosed with SCAN interviews (15) fulfilling a life-time, best estimate diagnosis of schizophrenia according to the ICD-10-DCR (16) and the DSM-IV (17). To minimise the effect of population stratification, recruitment was restricted to individuals of Danish ancestry for three generations. A sample of 350 healthy, psychiatrically unscreened Danish volunteer blood donors (controls) was obtained. In Denmark a health questionnaire must be completed and approved before blood donation. This ensures that none of the donors suffers from a current infectious disease. Due to restrictions defined by the ethical committees, ethnic origin is



Reference name	rs11003125	rs7096206	rs7095891	rs5030737	rs1800450	rs1800451
Relative position	- 550	- 221	4	223	230	239
Rare allele	H	X	Q	D	B	C
- freq. in controls	0.39	0.20	0.21	0.07	0.14	0.01
- freq. in patients	0.28	0.27	0.20	0.05	0.19	0.03

Fig. 1. Positions and frequencies of genetic markers in *MBL2*. Reference names and positions for the genetic markers in *MBL2* located at 10q21.1. The positions are relative to the untranslated (UTR) start position of exon 1. Allele frequencies of the rare alleles are given for controls and patients with schizophrenia.

unknown for the controls but they are expected to be mainly Western European descent. For the same reason no information on demographics is available. The studies were approved by the Danish Data Protection Agency and by the Danish Ethical Committees and the work has been carried out in accordance with the Helsinki Declaration.

DNA extraction and genotyping

Genomic DNA was extracted from whole blood using the Maxwell 16 System Blood DNA Purification Kit (Promega, Madison, WI, USA). Genotyping was performed using real-time polymerase chain reaction (rt-PCR) with TaqMan SNP Genotyping Assays (Applied Biosystems, Foster City, CA, USA).

MBL2−D (codon 52, rs5030737), *MBL2*−B (codon 54, rs1800450), *MBL2*−C (codon 57, rs1800451), *MBL2*−H/L (−550, rs11003125), *MBL2*−X/Y (−221, rs7096206) and *MBL2*−P/Q (+4, rs12780112) were genotyped using previously described assays (18−20). Genotyping for the *MASP2* mutation D120G [nucleotide 359 A to G (359 A/G)] was carried out in a similar way (19). The positions of the markers in *MBL2* are shown in Fig. 1.

For all TaqMan assays, DNA amplification was carried out in 384-well plates with 5 μl PCR containing 20 ng DNA, 0.9 μM primers and 0.2 μM probes (final concentrations). Reactions were performed with the following protocol on a GeneAmp PCR 9700: 95°C for 10 min, followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. To determine genotypes, endpoint fluorescence was read on the 7900 HT Sequence Detection Systems using SDS software version 2.3.

Haplotypes of the *MBL2* gene were identified (Table 1). Due to linkage disequilibrium, only seven haplotypes (HYPA, LYQA, LYPA, LXPA, LYPB, LYQC and HYPD) are commonly found, with HYPA being the most frequent in samples of European ancestry. In controls, nevertheless, an additional haplotype LYPD was found in a single individual. This sample was re-genotyped to exclude genotyping errors. LYPD has also been found in a few other studies (21−23). Recently Boldt et al. (24) explored the evolution of *MBL2* haplotypes and proposed a phylogenetic nomenclature to standardise studies related to *MBL2*. They suggest that LYPD probably is the product of a recent intragenic recombination event between HYPD and LYPA or LYPB. However, we excluded this individual rather than dealing with this extra haplotype. Two-marker haplotypes with mutant alleles (YB, YC and YD) were combined and collectively represented as YO, the other haplotypes being YA and XA. Genotypes based on these haplotypes were classified according to their known association with high (YA/YA, YA/XA), intermediate (XA/XA, YA/YO) or low/insufficient (XA/YO, YO/YO) MBL concentrations (25).

Table 1. *MBL2* haplotype and multilocus genotype frequencies: counts (proportions)

MBL2 haplotype	Controls (*N* = 349)	Patients (*N* = 100)
HYPA	223 (0.32)	47 (0.23)
LYPA	35 (0.05)	12 (0.06)
LY**Q**A	141 (0.20)	35 (0.17)
L**X**PA	139 (0.20)	53 (0.27)
HYP**D**	51 (0.07)	10 (0.05)
LYP**B**	100 (0.14)	38 (0.19)
LY**QC**	9 (0.01)	5 (0.03)
Total	698	200
Multilocus genotype		
High level of MBL		
YA/YA	119 (0.34)	17 (0.17)
YA/**X**A	77 (0.22)	36 (0.36)
Total	196 (0.56)	53 (0.53)
Intermediate		
XA/**X**A	15 (0.04)	4 (0.04)
YA/Y**O**	84 (0.24)	24 (0.24)
Total	99 (0.28)	28 (0.28)
Low/insufficient		
XA/Y**O**	32 (0.09)	9 (0.09)
Y**O**/Y**O**	22 (0.06)	10 (0.10)
Total	54 (0.15)	19 (0.19)

The rare alleles are marked with bold type and the O-allele is any of the D, B and C variants of exon 1. Multilocus genotypes are grouped with respect to their known association with high, intermediate or low/insufficient level of MBL in serum.

Concentration of MBL and MASP-2 in serum

Concentrations of MBL and MASP-2 in serum were determined as previously described (26). In brief, the method used was time resolved immunofluorometric assay (TRIFMA). For MBL assessment, serum samples, diluted 100-fold, were applied onto microtitre wells pre-coated with the polysaccharide, mannan, from baker's yeast. MBL binds through its carbohydrate-recognition domains and the bound MBL is detected with biotin-labelled monoclonal anti-MBL antibody, followed by europium-labelled streptavidin and time resolved fluorometry. MBL concentration was determined with a detection limit of 10 ng MBL/ml serum. Serum was missing for two of the patients.

MASP-2 concentration was also measured by TRIFMA (27). In brief, microtitre wells were coated with monoclonal anti-MASP-2 (MAb 8B5 against the C-terminal domains of MASP-2). Serum samples, diluted 40-fold, were applied, and bound MASP-2 was detected with biotin-labelled anti-MASP-2 (MAb 6G12 against the N-terminal domain of MASP-2), followed by europium-labelled streptavidin.

MBL deficiency classification is still an open question (28) and various serum levels have been suggested: <10, <50, <100 and <500 ng/ml. Often the detection limit of the specific assay has been used but no clinical data supports such a definition (29), and the clinical relevance may depend on the disease investigated (11). Only a minor part of deficient individuals become affected clinically. The following MBL levels will be referred to as: very low/deficient: <100, low: 100–400, normal: >400 ng/ml (http://www.ssi.dk).

Statistical analysis

Single-marker genotypic associations were assessed using logistic regression assuming an additive model on the log scale. The resulting odds ratio (OR) indicates the effect of each extra copy of the rare allele. Hence, the OR between the two homozygote variants is the square of the reported OR. Similarly, the additive effects of having 0, 1 or 2 copies for each of the seven haplotypes were considered. Linkage phase of haplotypes was assumed known although validity of the identified haplotypes was also checked by inferring phased haplotypes from genotypes with BEAGLE 2.1.3 (30). We ran BEAGLE 1000 times using a different seed (random starting point) for each run and observed that the multilocus genotypes matched perfectly in all runs (results not shown). Additive effects for each of m multiple SNPs were tested by an m df χ^2-test that has a corresponding score test which is a generalisation of the Armitage test (31).

The distributions of MBL and MASP-2 serum concentration were markedly skewed and clearly violate any assumption of a symmetric distribution (e.g. a normal distribution). Concentration of MBL and MASP-2 in serum were therefore analysed on log-transformed data. Standard analysis of variance (linear regression) was used for the analysis of MASP-2 concentration while Tobit regression analysis (32) was applied to handle the bulk of observations below the MBL 10 ng/ml detection limit (Fig. 2) by censoring techniques. A categorisation as indicated in the preceding subsection would also

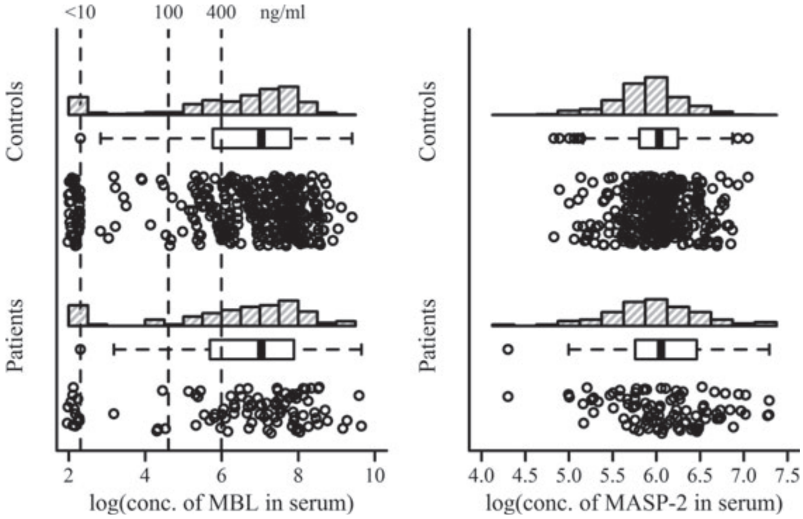


Fig. 2. Distribution of MBL and MASP-2 serum concentration. Concentration of MBL and MASP-2 in serum for controls and patients with schizophrenia. Before logarithmic transformation, concentrations were measured in ng protein/ml serum. The vertical lines in the left panel indicate: below MBL detection limit (<10 ng/ml), very low/insufficient MBL level (<100 ng/ml), low MBL level (100–400 ng/ml) and normal MBL level (>400 ng/ml). Histogram, box-plot and scatter plot of the observed concentrations are given for each protein and separately for patients and controls.

solve this problem but at the expense of continuity (information loss). Estimated median serum concentrations are presented after back-transformation with the exponential function.

Logistic regression was used for analyses of dichotomous traits of MBL deficiency status (</≥100 ng/ml) and MBL serum detection status (</≥10 ng/ml).

Statistical analyses were carried out using the software package R (http://www.r-project.org) and with a 5% level of significance. To account simultaneously for the nine different SNP and haplotype association tests permutation adjusted *p*-values were calculated using a step-down maximum-statistics approach corresponding to the algorithm from Box 2 in Dudoit et al.'s study (33). For serum concentration analyses solely nominal *p*-values were reported but these can be interpreted using the following Bonferroni thresholds: 0.01 for tests concerning MBL serum concentration and 0.0125 for tests concerning serum concentration of MASP-2. The following tests were primarily included to ease comparison with earlier studies and should be considered exploratory only in the context of multiple testing: two-marker haplotypes (YA/XA/YO), the A/O pseudo-marker and the tests of multiple SNPs.

Results

Haplotypes and multilocus genotypes

Allele frequencies of the genetic markers in *MBL2* are shown in Fig. 1. The most common mutation allele in exon 1 is B, while the D allele is common and the C allele is rare. With seven haplotypes there are 28 possible multilocus genotypes but 2 of these (LYPA/LYQC and LYQC/LYQC) were not observed.

Frequencies of haplotypes in the *MBL2* region are shown in Table 1. With 26 categories on only 100 and 350 individuals many of these will turn out having low expected counts and analyses using this 26 level variable would therefore be problematic. Grouping multilocus genotypes by X/Y and A/O has previously been used (34). Frequencies of the resulting six genotype groups are shown in Table 1. None were homozygous for the *MASP2* mutation allele 359 G/G. However, this is within expectations under Hardy–Weinberg proportions. The proportions of 359 A/G heterozygote individuals were 12% in patients and 9% in controls.

Association analysis

Results from the trend test of disease association with single- and multilocus genetic markers in *MBL2* are shown in Table 2. Significant association was found for the H/L marker and nominally for the X/Y marker. None of the SNPs in *MBL2* exon 1 were significantly associated with schizophrenia, possibly due to lack of power.

Table 2. Trend tests (1 df χ^2) for association of schizophrenia with *MBL2* single-locus and multilocus genetic markers by use of logistic regressions with an additive effect on a log scale of the rare allele (marked with bold type)

Locus	*p*-Value	OR (95% CI)	Adjusted *p*-value
Single			
H/L (m1)	0.0060	0.63 (0.45–0.88)	0.048
X/Y (m2)	0.047	1.46 (1.01–2.12)	0.25
P/**Q** (m3)	0.65	0.91 (0.61–1.34)	0.82
A/**D** (m4)	0.24	0.68 (0.32–1.29)	0.71
A/**B** (m5)	0.14	1.34 (0.90–1.95)	0.60
A/**C** (m6)	0.24	1.99 (0.60–5.90)	0.71
A/**O** (m7)*	0.32	1.19 (0.84–1.68)	—
Multi†			
HYPA	0.023	0.67 (0.47–0.95)	0.17
LYPA	0.59	1.20 (0.59–2.27)	0.82
LY**Q**A	0.39	0.84 (0.55–1.25)	0.72
YA‡	0.011	0.66 (0.48–0.91)	—

OR measure the effect of each extra copy of the rare allele and OR between the two homozygote variants is therefore this value squared. Permutation adjusted *p*-values from a step-down max-statistics procedure accounts simultaneously for the corresponding nine null hypotheses.

*The O-allele of the A/O marker is any of the D, B and C variants of *MBL2* exon 1.

†LXPA, HYPD, LYPB and LYQC are identifiable with m2, m4, m5 and m6, respectively.

‡XA and YO are identifiable with m2 and m7, respectively.

An exploratory analysis of multiple SNPs also showed nominal significant disease association with the three promoter region markers ($p = 0.016$) and for the X/Y marker combined with A/O ($p = 0.030$). It turns out that under the present conditions, the model for multiple SNPs with a trend parameter for each of the six single markers is simply another parameterisation of the model containing a trend parameter for each of the seven haplotypes (see Appendix S1, *Supporting Information*). The HYPA and YA haplotypes showed nominal significant protective effects against schizophrenia.

For *MASP2* no significant disease association was found although the proportion of 359 A/G heterozygotes was higher in patients.

Concentration of MBL and MASP-2 in serum

The median (and range) of the observed MBL concentration in serum for the three groups (high, intermediate and low/insufficient) determined by X/Y and A/O were: 2319.5 (202–12216), 446 (<10–1818) and <10 (<10–190) in patients; 2639 (557–15615), 538.5 (93–2558) and <10 (<10–444) in controls. The distribution of (log-transformed) MBL and MASP-2 serum concentrations are shown in Fig. 2.

Table 3 shows the results from Tobit regression analysis of MBL concentration in serum. As anticipated, significant single-marker and haplotype

Table 3. Association of MBL serum concentration with schizophrenia after adjustment for *MBL2* genetic variants (models 2, 3 and 4) as well as unadjusted (model 1)

MBL models	Parameters	Coefficient (95% CI)	Test statistic	*p*-Value
Model 1	Schizophrenia	−0.011 (−0.494 to 0.472)	−0.045	0.96
	Intercept	6.370 (6.144 to 6.597)		
Model 2*	Schizophrenia	0.213 (0.029 to 0.398)	2.26	0.024
	Intermediate	−1.672 (−1.845 to −1.500)		
	Low	−5.345 (−5.584 to −5.106)		
	Intercept	7.696 (7.588 to 7.803)		
Model 3†	Schizophrenia	0.412 (0.199 to 0.625)	3.79	1.5e-4
	XA	−1.318 (−1.478 to −1.159)		
	YO	−3.107 (−3.266 to −2.947)		
	Intercept	8.345 (8.201 to 8.489)		
Model 4	Schizophrenia	0.493 (0.291 to 0.694)	4.79	1.7e-6
	LYPA	−0.337 (−0.601 to −0.074)		
	LYQA	0.075 (−0.089 to 0.240)		
	LXPA	−1.300 (−1.464 to −1.135)		
	HYPD	−2.364 (−2.609 to −2.119)		
	LYPB	−3.388 (−3.579 to −3.197)		
	LYQC	−3.404 (−3.893 to −2.916)		
	Intercept	8.317 (8.128 to 8.506)		
Testing 4 vs. 3‡			57.5	9.6e-12

The additive effects of having 0, 1 or 2 copies for each of the specific haplotypes enter in the model 3 and 4.

*High level multilocus genotypes: YA/YA and YA/XA; intermediate: XA/XA and YA/YO; low: XA/YO and YO/YO.

†The O-allele of the A/O marker is any of the D, B and C variants of *MBL2* exon 1.

‡Deviance test (chi-square on 4 df) of the reduction from model 4 to model 3.

associations with MBL concentration were found with effects in the expected direction (results not shown). MBL concentrations in serum were not significantly different between patients and controls *per se* (model 1) with estimates of the median at 584 ng/ml [confidence interval (CI): 443–717] and 578 ng/ml (CI: 296–831), respectively (Table S1, *Supporting Information*). However, when adjusted for the additive effect of *MBL2* haplotypes the patient/control effect on MBL serum concentrations turned out being highly significant (model 4: $p =$ 1.7e-6) with a higher concentration in patients. Estimated median MBL concentrations in serum for each of the 26 observed multilocus genotypes are shown in Table S1. To ease comparison to other studies, we also show results obtained by use of the coarser X/Y-A/O groups.

The median (and range) of MASP-2 observed in patients and controls were 425 ng/ml (74–1467) and 417 ng/ml (125–1152), respectively. As expected, MASP-2 concentrations depended significantly on the *MASP2* genotypes (Table 4 model 2) and, surprisingly and interestingly, on *MBL2* genotypes (Table 4 model 4). Using a backward elimination procedure we found the effect of *MBL2* genotypes to be well captured by an additive effect of the O variant, i.e. number of alleles (0, 1 or 2) of any of the three mutations in *MBL2* exon 1 (results not shown).

Table 4. Regression analysis of (log-transformed) MASP-2 concentration in serum against the following independent variables: patients with schizophrenia versus controls, carrier of the rare *MASP2* G-allele (A/A and A/G) and number of *MBL2* exon 1 O-alleles (A/A: 0, A/O: 1 and O/O: 2) where the O-allele is any of the D, B and C variants of *MBL2* exon 1

MASP-2 models	Parameters	Coefficient (95% CI)	Test statistic*	*p*-Value
Model 1	Schizophrenia	0.071 (−0.023 to 0.165)	1.49	0.14
	Intercept	6.020 (5.976 to 6.064)		
Model 2	Schizophrenia	0.083 (−0.002 to 0.169)	1.92	0.055
	MASP2 A/G	−0.594 (−0.714 to −0.475)		
	Intercept	6.075 (6.034 to 6.116)		
Model 3	SZ : *MASP2* A/G	−0.354 (−0.628 to −0.080)	−2.54	0.011
	Schizophrenia (SZ)	0.121 (0.032 to 0.211)		
	MASP2 A/G	−0.505 (−0.642 to −0.368)		
	Intercept	6.067 (6.025 to 6.108)		
Model 4	*MBL2* O	0.115 (0.060 to 0.170)	4.12	4.46e-5
	SZ : *MASP2* A/G†	−0.358 (−0.627 to −0.088)		
	Schizophrenia (SZ)	0.112 (0.024 to 0.200)		
	MASP2 A/G	−0.502 (−0.637 to −0.367)		
	Intercept	6.014 (5.965 to 6.062)		

*Wald tests evaluated in a t-distribution with df equal to the difference between the number of subjects (349 + 98 = 447) and number of parameters in the model (e.g. 444 df in model 2).

†Here 'V_1: V_2' represents the interaction effect between V_1 and V_2.

MASP-2 concentrations in serum for patients and controls were not significantly different (Table 4). However, we found a significant interaction ($p =$ 0.011) between *MASP2* genotype and patient/control status (Table 4 model 3). Applying a forward inclusion procedure, the final model contained this interaction effect and also the additive effect of the *MBL2* variant (Table 4 model 4). Table S2 contains estimates of median MASP-2 serum concentrations from this final model and also results from the model without *MBL2* adjustment (Table 4 model 3) and with patient/control status only (Table 4 model 1). The 359 A/G variant was associated with lower MASP-2 concentrations in serum whereas mutations in *MBL2* exon 1 were associated with a higher level of MASP-2. The effect of the D120G mutation was stronger in patients than in controls (the interaction effect) and actually the difference between patients and controls within *MASP2* genotype changes direction. Specifically *MASP2* A/A patients have higher median MASP-2 concentration than the controls whereas this median is lower than controls for patients carrying the 359 A/G variant (Table S2).

MBL deficiency

Inherited MBL deficiency defined as homozygosity for either of the mutations in *MBL2* exon 1 (YO/YO in Table 1) was observed in a relatively high proportion (10%) of the patients although not significantly higher than in controls (6%). The proportion in controls is in line with the 5% detected from another Danish sample (35).

Distribution of MBL concentrations on the categories *very low* (<100), *low* (100–400) and *normal* (>400 ng/ml) is indicated in Fig. 2. With a 100 ng/ml cut-off (*very low*) 18% of the patients were deficient. Yet, this is not a significantly higher fraction than the 15% seen in controls.

Fifty-three individuals (12%) had MBL concentrations below the 10 ng/ml detection limit. This is usually seen in 10–15% of investigated individuals and is not a problem specific to this study. The proportion was not significantly different in patients (13%) and controls (11%). All O/O homozygous subjects except one belonged to this group. Amongst the 22 other of these 53 individuals, 21 (95%) carried the XA/YO genotype. All subjects with YA/YA had concentrations above the detection limit.

Discussion

Changes in inflammatory-related pathways have long been suggested to have a role in the pathophysiology of schizophrenia but there is no clear understanding as to which specific inflammatory-related pathways are involved or how they can precipitate the onset of the disorder. Activation of the peripheral innate immune cytokine pathways whether as a result of an immune challenge or stress leads to increased proinflammatory cytokine production and decreased neurotrophic support and neurogenesis in brain areas important to behaviour and cognition (36). In this study we attempt to link parameters of inflammation in the innate immune pathway with schizophrenia.

Constitutional, MBL levels of individuals with identical *MBL2* genotypes may vary 10-fold pointing to limitations in studies relying on genotyping only. Indeed, the lectin pathway of complement activation comprises several factors other than MBL. The associated serine proteases (MASP-1, MASP-2 and MASP-3) are thus required for the downstream transmission of the activation signal. Besides MBL, further three other recognition proteins, H-, L- and M-ficolins, may also initiate the lectin pathway. Apart from MBL our study included estimation of the main serine protease, MASP-2. It would have been satisfactory to include also MASP-1 in this study but to our knowledge nobody has yet been able to produce specific anti-MASP-1 antibodies. Thus, unfortunately, it is not possible to analyse for this component and as there is no assay, there is no literature on this. The involvement of the lectin pathway of the complement system was suggested by Mayilyan et al. (37). They found that the overall activity of the classical pathway as well as the C4 cleaving activity of MBL-MASP-2 complexes caught on a surface of mannan was elevated in patients with schizophrenia. Ideally, the remaining proteins in the pathway ought to be investigated and functional assessment would be relevant too (28). With regards to a more general measurement of complement factors in schizophrenia only very few studies have been performed and all with small sample sizes of less than 100 individuals (12).

The concentration of MBL was higher in patients when accounting for genetic variants of *MBL2* but the proportion of subjects in the MBL low-producing multilocus genotype groups was higher too, i.e. pointing toward increased risk of MBL deficiency indicating aetiological heterogeneity among patients. The genetic disease association was only significant for *MBL2* promoter region SNPs but the absence of association with the functional variants in exon 1 may be a power issue more than lacking true association. The frequency of the common HYPA haplotype was notable lower in patients (23%) as compared to controls (32%), in parallel with an increased frequency of especially LXPA and LYPB. The lower frequency of patients in the high level producing YA/YA group is largely compensated although by a higher frequency in the other high level producing group YA/XA. Therefore multilocus genotypes have to be considered when analysing for effects on MBL level. We investigated if the simplification by considering only X/Y and A/O markers was statistically justified by backward elimination from the saturated model but found this reduction to be significantly too coarse (results not shown). Thus, the detailed genotype grouping was preferred and in view of the estimates (Table S1) we recommend using the finer grid in future studies. Also, the recent work by Boldt et al. (24) should be taken into consideration.

A very clear effect of higher MBL serum concentration in patients with schizophrenia was seen when adjusting for the variation in MBL ascribed to *MBL2* variants. This may indicate that more insight into the aetiology of schizophrenia can be found by analysing the complement pathway of activation in greater detail. As schizophrenia is a polygenic disease (9) it is possible that there are variants in the genes coding for other complement components that are both associated with the disease and with elevated levels of MBL.

MASP-2 serum concentration was lower in patients than in controls for subjects carrying the D120G mutation but higher for wild-type carriers. This interaction was significant. In a sample of 492 Danes the allele frequency of the mutation was 3.6% but none were homozygous for the mutation and the clinical relevance of MASP-2 deficiency is uncertain (38). Interestingly, MASP-2 levels also depended significantly on variants in *MBL2* exon 1. This has not been reported before and we do not at present have a plausible explanation for this. It will be exciting to follow if other research groups find similar effects.

The findings from this study indicate an association between schizophrenia and components of the complement system, and future studies should explore the interplay between immunity-related genes in the human leukocyte antigen (HLA) region and key components in the lectin pathway. This would be of interest as the various individual MHC molecules encoded by the HLA region present a key factor of the adaptive immune system whereas the lectin pathway represents the innate immune system. The recent study by Håvik et al. (6) seconds this with findings of significant association between schizophrenia and immunity-related genes both within and outside the HLA region.

In conclusion this study supports previous studies showing increased complement activity in patients with schizophrenia, it indicates aetiological heterogeneity among patients and underline that multilocus genotypes have to be considered when the effect on MBL level is investigated. It is apparent that inclusion of additional components from the complement system will be vital to investigate further the association between schizophrenia and the activation pathway.

Acknowledgements

We thank Annette G. Hansen and Lisbeth Jensen for technical assistance, Jakob Grove for a critical evaluation of the manuscript and Raben Rosenberg for valuable discussions and comments. Funding for this study was provided by the Danish Medical Research Foundation (DMRF) and The Stanley Medical Research Institute. The authors declare no conflicts of interests.

References

1. Yolken RH, Torrey EF. Viruses, schizophrenia, and bipolar disorder. Clin Microbiol Rev 1995;**8**:131–145.
2. Eaton WW, Byrne M, Ewald H et al. Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. Am J Psychiatry 2006;**163**:521–528.
3. Buka SL, Tsuang MT, Torrey EF, Klebanoff MA, Bernstein D, Yolken RH. Maternal infections and subsequent psychosis among offspring. Arch Gen Psychiatry 2001;**58**:1032–1037.
4. Yolken RH, Torrey EF. Are some cases of psychosis caused by microbial agents? A review of the evidence. Mol Psychiatry 2008;**13**:470–479.
5. Xiao JC, Buka SL, Cannon TD et al. Serological pattern consistent with infection with type I Toxoplasma gondii in mothers and risk of psychosis among adult offspring. Microbes Infect 2009;**11**:1011–1018.
6. Håvik B, Hellard SL, Rietschel M et al. The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia. Biol Psychiatry 2011;**70**:35–42.
7. Kraus DM, Elliott GS, Chute H et al. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. J Immunol 2006;**176**:4419–4430.
8. Stefansson H, Ophoff RA, Steinberg S et al. Common variants conferring risk of schizophrenia. Nature 2009;**460**: 744–747.
9. International Schizophrenia Consortium, Wray NR, Stone JL et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;**460**: 748–752.
10. Shi J, Levinson DF, Duan J et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 2009;**460**:753–757.
11. Thiel S, Frederiksen PD, Jensenius JC. Clinical manifestations of mannan-binding lectin deficiency. Mol Immunol 2006;**43**:86–93.
12. Mayilyan KR, Weinberger DR, Sim RB. The complement system in schizophrenia. Drug News Perspect 2008;**21**: 200–210.
13. Garred P, Honore C, Ma YJ, Munthe-Fog L, Hummelshøj T. MBL2, FCN1, FCN2 and FCN3-The genes behind the initiation of the lectin pathway of complement. Mol Immunol 2009;**46**:2737–2744.
14. Garred P, Larsen F, Seyfarth J, Fujita R, Madsen HO. Mannose-binding lectin and its genetic variants. Genes Immun 2006;**7**:85–94.
15. Wing JK, Sartorius N, Üstün TB. Diagnosis and clinical measurement in psychiatry. A reference manual for SCAN. Cambridge: Cambridge University Press, 1998.
16. World Health Organization. The ICD-10 classification of mental and behavioural disorders. Diagnostic criteria for research. Geneva: World Health Organization, 1993.
17. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. DSM-IV, 4th edn. Washington, DC: American Psychiatric Association, 1994.
18. Henckaerts L, Nielsen KR, Steffensen R et al. Polymorphisms in innate immunity genes predispose to bacteremia and death in the medical intensive care unit. Crit Care Med 2009;**37**:192–201.
19. Mølle I, Steffensen R, Thiel S, Peterslund NA. Chemotherapy-related infections in patients with multiple myeloma: associations with mannan-binding lectin genotypes. Eur J Haematol 2006;**77**:19–26.
20. Van HE, Houtmeyers F, Massonet C et al. Detection of single nucleotide polymorphisms in the mannose-binding lectin gene using minor groove binder-DNA probes. J Immunol Methods 2004;**287**:227–230.
21. Boldt AB, Petzl-Erler ML. A new strategy for mannose-binding lectin gene haplotyping. Hum Mutat 2002;**19**: 296–306.
22. Cedzynski M, Szemraj J, Swierzko AS et al. Mannan-binding lectin insufficiency in children with recurrent infections of the respiratory system. Clin Exp Immunol 2004;**136**: 304–311.
23. Skalnikova H, Freiberger T, Chumchalova J, Grombirikova H, Sediva A. Cost-effective genotyping of human MBL2 gene mutations using multiplex PCR. J Immunol Methods 2004;**295**:139–147.
24. Boldt AB, Messias-Reason IJ, Meyer D et al. Phylogenetic nomenclature and evolution of mannose-binding lectin (MBL2) haplotypes. BMC Genet 2010;**11**:38.
25. Olesen HV, Jensenius JC, Steffensen R, Thiel S, Schiøtz PO. The mannan-binding lectin pathway and lung disease in cystic fibrosis – dysfunction of mannan-binding

lectin-associated serine protease 2 (MASP-2) may be a major modifier. Clin Immunol 2006;**121**:324–331.

26. Thiel S, Møller-Kristensen M, Jensen L, Jensenius JC. Assays for the functional activity of the mannan-binding lectin pathway of complement activation. Immunobiology 2002;**205**:446–454.
27. Møller-Kristensen M, Jensenius JC, Jensen L et al. Levels of mannan-binding lectin-associated serine protease-2 in healthy individuals. J Immunol Methods 2003;**282**: 159–167.
28. Dommett RM, Klein N, Turner MW. Mannose-binding lectin in innate immunity: past, present and future. Tissue Antigens 2006;**68**:193–209.
29. Petersen SV, Thiel S, Jensenius JC. The mannan-binding lectin pathway of complement activation: biology and disease association. Mol Immunol 2001;**38**:133–149.
30. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;**81**:1084–1097.
31. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;**7**:781–791.
32. Amemiya T. Tobit models - a Survey. J Econom 1984;**24**: 3–61.
33. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003;**18**:71–103.
34. Steffensen R, Hoffmann K, Varming K. Rapid genotyping of MBL2 gene mutations using real-time PCR with fluorescent hybridisation probes. J Immunol Methods 2003;**278**: 191–199.
35. Dahl M, Tybjærg-Hansen A, Schnohr P, Nordestgaard BG. A population-based study of morbidity and mortality in mannose-binding lectin deficiency. J Exp Med 2004;**199**:1391–1399.
36. Capuron L, Miller AH. Immune system to brain signaling: neuropsychopharmacological implications. Pharmacol Ther 2011;**130**:266–238.
37. Mayilyan KR, Arnold JN, Presanis JS, Soghoyan AF, Sim RB. Increased complement classical and mannan-binding lectin pathway activities in schizophrenia. Neurosci Lett 2006;**404**:336–341.
38. Sørensen R, Thiel S. Jensenius JC. Mannan-binding-lectin-associated serine proteases, characteristics and disease associations. Springer Semin Immunopathol 2005;**27**: 299–319.

Supporting Information

The following Supporting information is available for this article:

Appendix S1. Methods.

Table S1. Estimated median MBL concentration in serum.

Table S2. Estimated median MASP-2 concentration in serum.

Additional Supporting information may be found in the online version of this article.

Please note: Wiley-Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

6.1.1 Paper 1 - supplementary info

Supplementary Methods

In the present study we observe seven haplotypes HYPA, LYPA, LYQA, LXPA, HYPD, LYPB and LYQC from the six SNPs H/L, X/Y, P/Q, A/D, A/B and A/C. We will refer to these six SNPs by their rare allele: H, X, Q, D, B and C, respectively. The latter three are positioned in *MBL2* exon 1 and never occurs in the same chromosome, i.e. these can be seen as one 4-allelic marker and explains why the haplotypes can be represented by 4 alleles. Formally these would be HYPAAA, LYPAAA, LYQAAA, LXPAAA, HYPDAA, LYPABA and LYQAAC. We will code genotypes and multilocus genotypes by the number of rare alleles, i.e. 0 (C/C), 1 (C/R) and 2 (R/R) where C and R denotes the common and rare allele, respectively. Obviously, each individual carry exactly two of these haplotypes: 0, 1 or 2 of each. Thus, with the 0/1/2 coding the sum of the seven possible haplotypes will be two for each individual. Clearly, the following set of seven equations is therefore true:

$$
\begin{aligned}
&H = HYPA + HYPD = 2 - (LYPA + LYQA + LXPA + LYPB + LYQC) = 2 - L \\
&X = LXPA \\
&Q = LYQA + LYQC \\
&D = HYPD \\
&B = LYPB \\
&C = LYQC \\
&2 = HYPA + LYPA + LYQA + LXPA + HYPD + LYPB + LYQC
\end{aligned}
$$

A direct calculation now gives that the model for multiple SNPs with a trend parameter for each of the six single markers, logit(p) = $\alpha_0 + \alpha_1 H + \alpha_2 X + \alpha_3 Q + \alpha_4 D + \alpha_5 B + \alpha_6 C$, can be re-parameterised to the model containing a trend parameter for each of the seven haplotypes, logit(p) = $\beta_0 + \beta_1 LYPA + \beta_2 LYQA + \beta_3 LXPA + \beta_4 HYPD + \beta_5 LYPB + \beta_6 LYQC$, with $\beta_0 = \alpha_0 + 2\alpha_1$, $\beta_1 = -\alpha_1$, $\beta_2 = \alpha_3 - \alpha_1$, $\beta_3 = \alpha_2 - \alpha_1$, $\beta_4 = \alpha_4$, $\beta_5 = \alpha_5 - \alpha_1$, and $\beta_6 = \alpha_6 + \alpha_3 - \alpha_1$.

Supplementary Tables

Supplementary table legends

Supplementary Table 1. Estimated median MBL concentration in serum.

Median MBL concentration in serum (ng/ml) estimated from Tobit regressions (on log-transformed data) with patients and controls specific means and an additive effect of haplotypes (Table 3 model 4). The 95% confidence intervals in the parentheses were estimated by use of the normal approximation on results from ordinary bootstrapping with 10000 replicates (Davison, A.C., Hinkley, D.V., Canty, A.J., Bootstrap methods and their application, Cambridge University Press, Cambridge, 1997). Results from using only the two markers X/Y and A/O (Table 3 model 3) are given before the corresponding four-marker multilocus genotype groups (e.g. YA/YA corresponds to YA=2, XA=YO=0). The results in the first row (Any) are from the model with only patient/control status as a factor (Table 3 model 1).

Supplementary Table 2. Estimated median MASP-2 concentration in serum.

Median MASP-2 concentration in serum (ng/ml) estimated from regressions analysis (on log-transformed data) with patients, controls and *MASP2* genotype specific means (interaction effect) and a linear effect of the O variant (A/O) in *MBL2* exon 1 (Table 4 model 4). The 95% confidence intervals (in the parentheses) were estimated by use of the normal approximation on results from ordinary bootstrapping with 10000 replicates. Results obtained from Table 4 model 3 (patient/control; *MASP2* genotype and the interaction effect) are given before the corresponding combination with *MBL2* genotype (*MASP2* A/A and *MASP2* A/G rows)). The results in the first row (Any) are from Table 4 model 1 which only includes the patient/control factor (i.e. any genotype combination).

Supplementary Table 1. Estimated median MBL concentration in serum.

***MBL2* genotype**	**Controls (N = 349)**	**Patients (N = 98)**
Any	584 (443 - 717)	578 (296 - 831)
YA/YA	4210 (3649 - 4754)	6357 (4907 - 7729)
HYPA/HYPA	4092 (3438 - 4733)	6698 (5017 - 8277)
HYPA/LYPA	2920 (1819 - 3923)	4780 (2897 - 6474)
HYPA/LYQA	4412 (3863 - 4961)	7221 (5514 - 8836)
LYPA/LYPA	2084 (413 - 3485)	3411 (723 - 5658)
LYQA/LYPA	3148 (1965 - 4229)	5153 (3103 - 7000)
LYQA/LYQA	4757 (3553 - 5731)	7785 (5437 - 9979)
YA/XA	1126 (990 - 1261)	1701 (1370 - 2019)
HYPA/LXPA	1116 (974 - 1253)	1826 (1449 - 2179)
LYPA/LXPA	796 (493 - 1069)	1303 (805 - 1749)
LYQA/LXPA	1203 (1 030 - 1370)	1969 (1523 - 2386)
XA/XA	301 (221 - 378)	455 (322 - 580)
LXPA/LXPA	304 (228 - 375)	498 (357 - 627)
YA/YO	188 (156 - 219)	285 (221 - 344)
HYPA/LYPB	138 (109 - 165)	226 (169 - 279)
LYPA/LYPB	99 (56 - 136)	161 (93 - 222)
LYQA/LYPB	149 (117 - 179)	244 (180 - 303)
HYPA/LYQC	136 (31 - 226)	223 (43 - 375)
LYQA/LYQC	147 (35 - 243)	240 (47 - 405)
HYPA/HYPD	385 (259 - 498)	630 (380 - 851)
LYPA /HYPD	275 (139 - 393)	449 (216 - 650)
LYQA /HYPD	415 (278 - 538)	679 (405 - 922)
XA/YO	50 (38 - 62)	76 (56 - 95)
LXPA/HYPD	105 (68 - 138)	172 (102 - 233)
LXPA/LYPB	38 (28 - 47)	62 (44 - 78)
LXPA/LYQC	37 (7 - 63)	61 (10 - 104)
YO/YO	8 (5 - 11)	13 (8 - 17)
HYPD/HYPD	36 (10 - 58)	59 (13 - 97)
HYPD/LYPB	13 (8 - 18)	21 (12 - 29)
HYPD/LYQC	13 (2 - 22)	21 (2 - 37)
LYPB/LYPB	5 (3 - 6)	8 (4 - 10)
LYPB/LYQC	5 (1 - 8)	8 (1 - 13)

Supplementary Table 2. Estimated median MASP-2 concentration in serum.

MBL2 genotype	Controls (N = 349)	Patients (N = 98)
Any	412 (395 - 428)	442 (394 - 488)
***MASP2* A/A**	431 (415 - 448)	487 (438 - 533)
A/A	409 (390 - 428)	457 (409 - 503)
A/O	459 (437 - 480)	513 (461 - 563)
O/O	515 (466 - 563)	576 (502 - 648)
***MASP2* A/G**	260 (223 - 295)	206 (146 - 261)
A/A	247 (214 - 279)	194 (139 - 245)
A/O	278 (239 - 315)	217 (157 - 274)
O/O	312 (259 - 362)	244 (173 - 309)

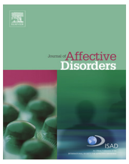
6.2 Paper 2[31]

Journal of Affective Disorders 164 (2014) 148–154

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad



Research report

Bipolar and panic disorders may be associated with hereditary defects in the innate immune system

Leslie Foldager[a,b,c,*], Ole Kohler[c,d], Rudi Steffensen[e], Steffen Thiel[f], Ann Suhl Kristensen[g], Jens Christian Jensenius[f], Ole Mors[c,d]

[a] Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[b] Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[c] iPSYCH: The Lundbeck Foundation Initiative for Integrated Psychiatric Research, Aarhus and Copenhagen, Denmark
[d] Research Department P, Aarhus University Hospital, Risskov, Denmark
[e] Department of Clinical Immunology, Aalborg University Hospital, Aalborg, Denmark
[f] Department of Biomedicine, Aarhus University, Aarhus, Denmark
[g] Regional Psychiatric Services West, Central Denmark Region, Herning, Denmark

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form
9 April 2014
Accepted 10 April 2014
Available online 19 April 2014

Keywords:
Mannan-binding lectin (MBL)
Innate immune system
Panic disorder
Bipolar disorder
Mannan-binding lectin-associated serine protease-2 (MASP-2)

ABSTRACT

Background: Mannan-binding lectin (MBL) and mannan-binding lectin-associated serine protease-2 (MASP-2) represent important arms of the innate immune system, and different deficiencies may result in infections or autoimmune diseases. Both bipolar and panic disorders are associated with increased inflammatory response, infections and mutual comorbidity. However, associations with MBL, MASP-2 or the gene, *MBL2*, coding for MBL, have not been investigated thoroughly.
Methods: One hundred patients with bipolar disorder, 100 with panic disorder and 349 controls were included. Serum concentrations of MBL and MASP-2 were measured and seven single nucleotide polymorphisms (SNPs) influencing these concentrations were genotyped. Disease association with genetic markers and serum levels were investigated.
Results: In panic disorder, we observed a large proportion (30%) of MBL deficient (< 100 ng/ml) individuals and significantly lower levels of MBL and MASP-2 plus association with the *MBL2* YA two-marker haplotype. Bipolar disorder was associated with the *MBL2* LXPA haplotype and lower MASP-2 levels.
Limitations: No information on course or severity of disorders was included, and only MBL and MASP-2 were measured, excluding other components from the complement pathway. Restrictions defined by ethnical committees preclude information of control's ethnic origin.
Conclusions: Significant differences in MBL and MASP-2 concentrations were observed between cohorts, especially an intriguing finding associating panic disorder with MBL deficiency. These differences could not be fully explained by allele or haplotype frequency variations. Since MBL deficiency is highly heterogeneous and associated with both infectious and autoimmune states, more research is needed to identify which complement pathway components could be associated with bipolar respectively panic disorder.

1. Introduction

Mannan-binding lectin (MBL) deficiency is the most common hereditary defect in the human innate immune system with an estimated prevalence of 10–30%, depending on definition (Thiel et al., 2006). The heterogeneity of MBL deficiency is emphasized in recent studies, since it may increase the susceptibility for both infections and autoimmune states (Heitzeneder et al., 2012; Mayilyan, 2012).

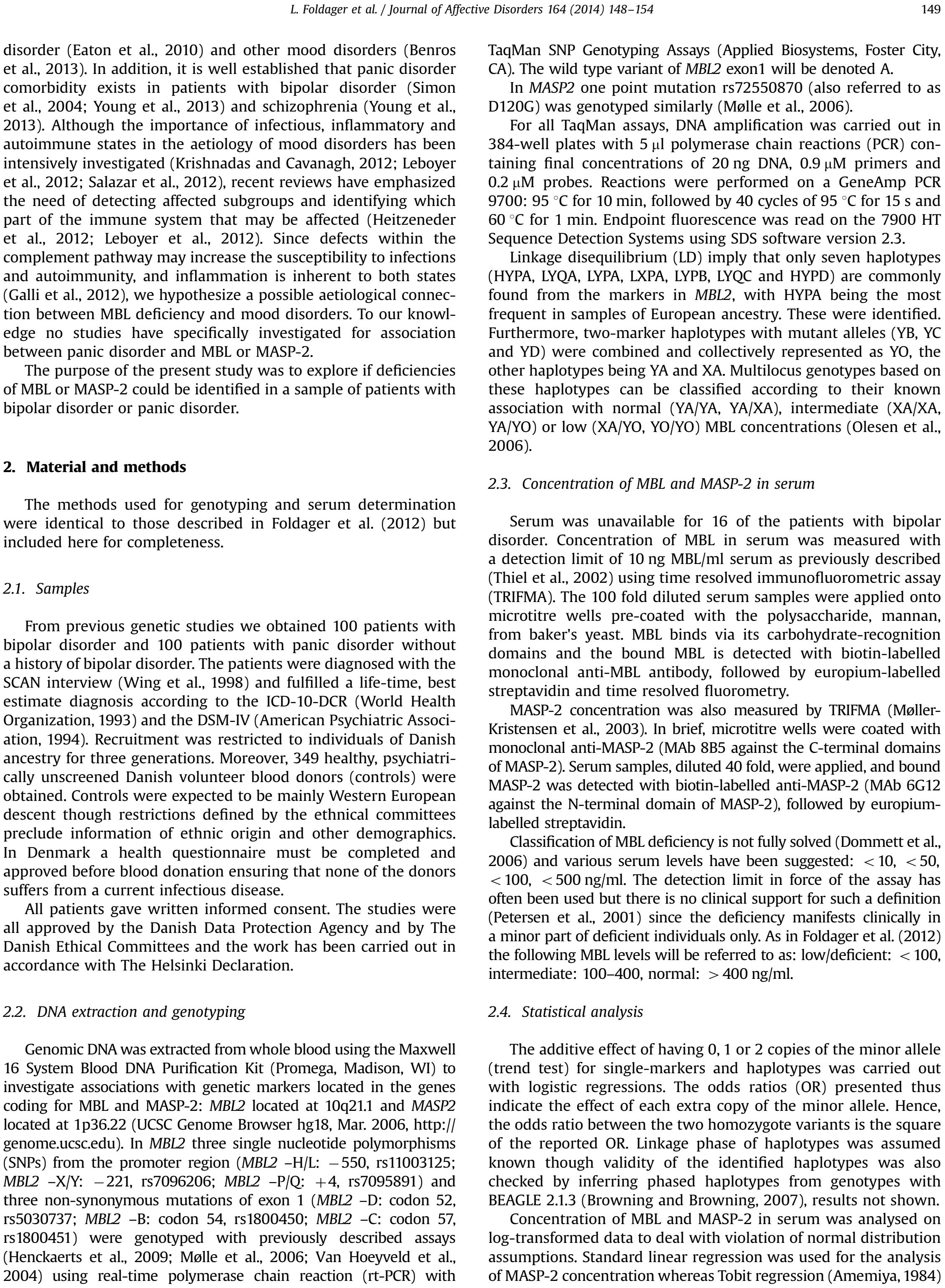
Several lines of evidence suggest that infection, inflammation and autoimmunity may contribute to the aetiology of schizophrenia (Benros et al., 2011, 2012; Fillman et al., 2013), and in a recent study, we associated schizophrenia with MBL and mannan-binding lectin-associated serine protease-2 (MASP-2), two key components of the lectin pathway of complement activation (Foldager et al., 2012).

The aetiology of both bipolar (Leboyer et al., 2012) and panic respectively anxiety disorder (Salazar et al., 2012; Chen et al., 2013) is suspected to be associated with an inflammatory state, and autoimmune processes and infections may precede bipolar

* Corresponding author at: Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Skovagervej 2, DK8240 Risskov, Denmark. Tel.: +45 78471119; fax: +45 78471108.
E-mail address: Leslie@birc.au.dk (L. Foldager).

http://dx.doi.org/10.1016/j.jad.2014.04.017
0165-0327/© 2014 Elsevier B.V. All rights reserved.

[31] Journal of Affective Disorders 2014; **164**: 148–154. © Elsevier. Reused with permission.



disorder (Eaton et al., 2010) and other mood disorders (Benros et al., 2013). In addition, it is well established that panic disorder comorbidity exists in patients with bipolar disorder (Simon et al., 2004; Young et al., 2013) and schizophrenia (Young et al., 2013). Although the importance of infectious, inflammatory and autoimmune states in the aetiology of mood disorders has been intensively investigated (Krishnadas and Cavanagh, 2012; Leboyer et al., 2012; Salazar et al., 2012), recent reviews have emphasized the need of detecting affected subgroups and identifying which part of the immune system that may be affected (Heitzeneder et al., 2012; Leboyer et al., 2012). Since defects within the complement pathway may increase the susceptibility to infections and autoimmunity, and inflammation is inherent to both states (Galli et al., 2012), we hypothesize a possible aetiological connection between MBL deficiency and mood disorders. To our knowledge no studies have specifically investigated for association between panic disorder and MBL or MASP-2.

The purpose of the present study was to explore if deficiencies of MBL or MASP-2 could be identified in a sample of patients with bipolar disorder or panic disorder.

2. Material and methods

The methods used for genotyping and serum determination were identical to those described in Foldager et al. (2012) but included here for completeness.

2.1. Samples

From previous genetic studies we obtained 100 patients with bipolar disorder and 100 patients with panic disorder without a history of bipolar disorder. The patients were diagnosed with the SCAN interview (Wing et al., 1998) and fulfilled a life-time, best estimate diagnosis according to the ICD-10-DCR (World Health Organization, 1993) and the DSM-IV (American Psychiatric Association, 1994). Recruitment was restricted to individuals of Danish ancestry for three generations. Moreover, 349 healthy, psychiatrically unscreened Danish volunteer blood donors (controls) were obtained. Controls were expected to be mainly Western European descent though restrictions defined by the ethnical committees preclude information of ethnic origin and other demographics. In Denmark a health questionnaire must be completed and approved before blood donation ensuring that none of the donors suffers from a current infectious disease.

All patients gave written informed consent. The studies were all approved by the Danish Data Protection Agency and by The Danish Ethical Committees and the work has been carried out in accordance with The Helsinki Declaration.

2.2. DNA extraction and genotyping

Genomic DNA was extracted from whole blood using the Maxwell 16 System Blood DNA Purification Kit (Promega, Madison, WI) to investigate associations with genetic markers located in the genes coding for MBL and MASP-2: *MBL2* located at 10q21.1 and *MASP2* located at 1p36.22 (UCSC Genome Browser hg18, Mar. 2006, http://genome.ucsc.edu). In *MBL2* three single nucleotide polymorphisms (SNPs) from the promoter region (*MBL2* –H/L: −550, rs11003125; *MBL2* –X/Y: −221, rs7096206; *MBL2* –P/Q: +4, rs7095891) and three non-synonymous mutations of exon 1 (*MBL2* –D: codon 52, rs5030737; *MBL2* –B: codon 54, rs1800450; *MBL2* –C: codon 57, rs1800451) were genotyped with previously described assays (Henckaerts et al., 2009; Mølle et al., 2006; Van Hoeyveld et al., 2004) using real-time polymerase chain reaction (rt-PCR) with TaqMan SNP Genotyping Assays (Applied Biosystems, Foster City, CA). The wild type variant of *MBL2* exon1 will be denoted A.

In *MASP2* one point mutation rs72550870 (also referred to as D120G) was genotyped similarly (Mølle et al., 2006).

For all TaqMan assays, DNA amplification was carried out in 384-well plates with 5 μl polymerase chain reactions (PCR) containing final concentrations of 20 ng DNA, 0.9 μM primers and 0.2 μM probes. Reactions were performed on a GeneAmp PCR 9700: 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Endpoint fluorescence was read on the 7900 HT Sequence Detection Systems using SDS software version 2.3.

Linkage disequilibrium (LD) imply that only seven haplotypes (HYPA, LYQA, LYPA, LXPA, LYPB, LYQC and HYPD) are commonly found from the markers in *MBL2*, with HYPA being the most frequent in samples of European ancestry. These were identified. Furthermore, two-marker haplotypes with mutant alleles (YB, YC and YD) were combined and collectively represented as YO, the other haplotypes being YA and XA. Multilocus genotypes based on these haplotypes can be classified according to their known association with normal (YA/YA, YA/XA), intermediate (XA/XA, YA/YO) or low (XA/YO, YO/YO) MBL concentrations (Olesen et al., 2006).

2.3. Concentration of MBL and MASP-2 in serum

Serum was unavailable for 16 of the patients with bipolar disorder. Concentration of MBL in serum was measured with a detection limit of 10 ng MBL/ml serum as previously described (Thiel et al., 2002) using time resolved immunofluorometric assay (TRIFMA). The 100 fold diluted serum samples were applied onto microtitre wells pre-coated with the polysaccharide, mannan, from baker's yeast. MBL binds via its carbohydrate-recognition domains and the bound MBL is detected with biotin-labelled monoclonal anti-MBL antibody, followed by europium-labelled streptavidin and time resolved fluorometry.

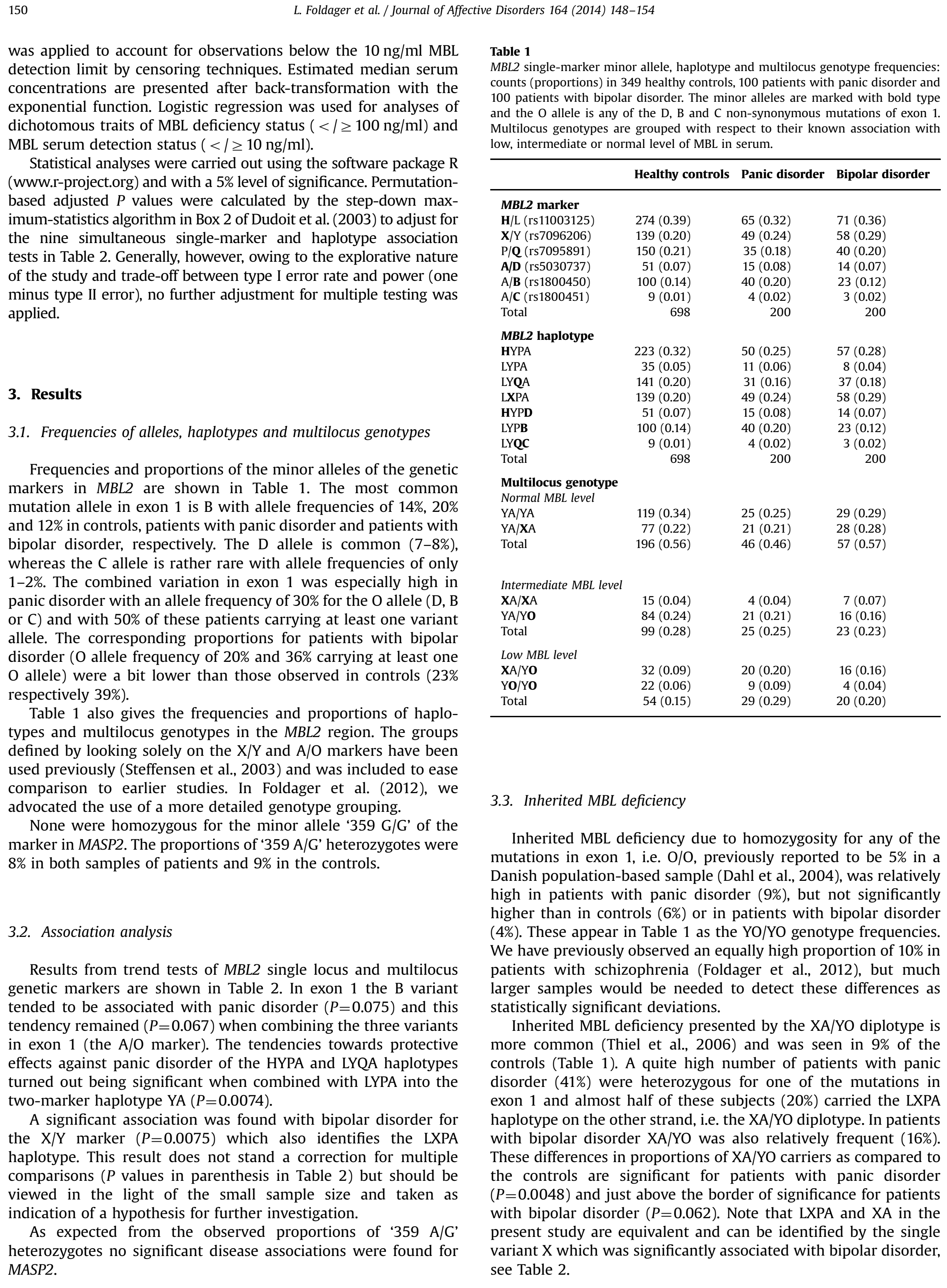
MASP-2 concentration was also measured by TRIFMA (Møller-Kristensen et al., 2003). In brief, microtitre wells were coated with monoclonal anti-MASP-2 (MAb 8B5 against the C-terminal domains of MASP-2). Serum samples, diluted 40 fold, were applied, and bound MASP-2 was detected with biotin-labelled anti-MASP-2 (MAb 6G12 against the N-terminal domain of MASP-2), followed by europium-labelled streptavidin.

Classification of MBL deficiency is not fully solved (Dommett et al., 2006) and various serum levels have been suggested: <10, <50, <100, <500 ng/ml. The detection limit in force of the assay has often been used but there is no clinical support for such a definition (Petersen et al., 2001) since the deficiency manifests clinically in a minor part of deficient individuals only. As in Foldager et al. (2012) the following MBL levels will be referred to as: low/deficient: <100, intermediate: 100–400, normal: >400 ng/ml.

2.4. Statistical analysis

The additive effect of having 0, 1 or 2 copies of the minor allele (trend test) for single-markers and haplotypes was carried out with logistic regressions. The odds ratios (OR) presented thus indicate the effect of each extra copy of the minor allele. Hence, the odds ratio between the two homozygote variants is the square of the reported OR. Linkage phase of haplotypes was assumed known though validity of the identified haplotypes was also checked by inferring phased haplotypes from genotypes with BEAGLE 2.1.3 (Browning and Browning, 2007), results not shown.

Concentration of MBL and MASP-2 in serum was analysed on log-transformed data to deal with violation of normal distribution assumptions. Standard linear regression was used for the analysis of MASP-2 concentration whereas Tobit regression (Amemiya, 1984)



was applied to account for observations below the 10 ng/ml MBL detection limit by censoring techniques. Estimated median serum concentrations are presented after back-transformation with the exponential function. Logistic regression was used for analyses of dichotomous traits of MBL deficiency status ($<$ / $\geq$ 100 ng/ml) and MBL serum detection status ($<$ / $\geq$ 10 ng/ml).

Statistical analyses were carried out using the software package R (www.r-project.org) and with a 5% level of significance. Permutation-based adjusted *P* values were calculated by the step-down max-imum-statistics algorithm in Box 2 of Dudoit et al. (2003) to adjust for the nine simultaneous single-marker and haplotype association tests in Table 2. Generally, however, owing to the explorative nature of the study and trade-off between type I error rate and power (one minus type II error), no further adjustment for multiple testing was applied.

3. Results

3.1. Frequencies of alleles, haplotypes and multilocus genotypes

Frequencies and proportions of the minor alleles of the genetic markers in *MBL2* are shown in Table 1. The most common mutation allele in exon 1 is B with allele frequencies of 14%, 20% and 12% in controls, patients with panic disorder and patients with bipolar disorder, respectively. The D allele is common (7–8%), whereas the C allele is rather rare with allele frequencies of only 1–2%. The combined variation in exon 1 was especially high in panic disorder with an allele frequency of 30% for the O allele (D, B or C) and with 50% of these patients carrying at least one variant allele. The corresponding proportions for patients with bipolar disorder (O allele frequency of 20% and 36% carrying at least one O allele) were a bit lower than those observed in controls (23% respectively 39%).

Table 1 also gives the frequencies and proportions of haplotypes and multilocus genotypes in the *MBL2* region. The groups defined by looking solely on the X/Y and A/O markers have been used previously (Steffensen et al., 2003) and was included to ease comparison to earlier studies. In Foldager et al. (2012), we advocated the use of a more detailed genotype grouping.

None were homozygous for the minor allele '359 G/G' of the marker in *MASP2*. The proportions of '359 A/G' heterozygotes were 8% in both samples of patients and 9% in the controls.

3.2. Association analysis

Results from trend tests of *MBL2* single locus and multilocus genetic markers are shown in Table 2. In exon 1 the B variant tended to be associated with panic disorder ($P=0.075$) and this tendency remained ($P=0.067$) when combining the three variants in exon 1 (the A/O marker). The tendencies towards protective effects against panic disorder of the HYPA and LYQA haplotypes turned out being significant when combined with LYPA into the two-marker haplotype YA ($P=0.0074$).

A significant association was found with bipolar disorder for the X/Y marker ($P=0.0075$) which also identifies the LXPA haplotype. This result does not stand a correction for multiple comparisons (*P* values in parenthesis in Table 2) but should be viewed in the light of the small sample size and taken as indication of a hypothesis for further investigation.

As expected from the observed proportions of '359 A/G' heterozygotes no significant disease associations were found for *MASP2*.

Table 1
MBL2 single-marker minor allele, haplotype and multilocus genotype frequencies: counts (proportions) in 349 healthy controls, 100 patients with panic disorder and 100 patients with bipolar disorder. The minor alleles are marked with bold type and the O allele is any of the D, B and C non-synonymous mutations of exon 1. Multilocus genotypes are grouped with respect to their known association with low, intermediate or normal level of MBL in serum.

	Healthy controls	Panic disorder	Bipolar disorder
***MBL2* marker**			
H/L (rs11003125)	274 (0.39)	65 (0.32)	71 (0.36)
X/Y (rs7096206)	139 (0.20)	49 (0.24)	58 (0.29)
P/**Q** (rs7095891)	150 (0.21)	35 (0.18)	40 (0.20)
A/D (rs5030737)	51 (0.07)	15 (0.08)	14 (0.07)
A/**B** (rs1800450)	100 (0.14)	40 (0.20)	23 (0.12)
A/**C** (rs1800451)	9 (0.01)	4 (0.02)	3 (0.02)
Total	698	200	200
***MBL2* haplotype**			
HYPA	223 (0.32)	50 (0.25)	57 (0.28)
LYPA	35 (0.05)	11 (0.06)	8 (0.04)
LY**Q**A	141 (0.20)	31 (0.16)	37 (0.18)
L**X**PA	139 (0.20)	49 (0.24)	58 (0.29)
HYP**D**	51 (0.07)	15 (0.08)	14 (0.07)
LYP**B**	100 (0.14)	40 (0.20)	23 (0.12)
LY**QC**	9 (0.01)	4 (0.02)	3 (0.02)
Total	698	200	200
Multilocus genotype			
Normal MBL level			
YA/YA	119 (0.34)	25 (0.25)	29 (0.29)
YA/**X**A	77 (0.22)	21 (0.21)	28 (0.28)
Total	196 (0.56)	46 (0.46)	57 (0.57)
Intermediate MBL level			
XA/**X**A	15 (0.04)	4 (0.04)	7 (0.07)
YA/Y**O**	84 (0.24)	21 (0.21)	16 (0.16)
Total	99 (0.28)	25 (0.25)	23 (0.23)
Low MBL level			
XA/Y**O**	32 (0.09)	20 (0.20)	16 (0.16)
Y**O**/Y**O**	22 (0.06)	9 (0.09)	4 (0.04)
Total	54 (0.15)	29 (0.29)	20 (0.20)

3.3. Inherited MBL deficiency

Inherited MBL deficiency due to homozygosity for any of the mutations in exon 1, i.e. O/O, previously reported to be 5% in a Danish population-based sample (Dahl et al., 2004), was relatively high in patients with panic disorder (9%), but not significantly higher than in controls (6%) or in patients with bipolar disorder (4%). These appear in Table 1 as the YO/YO genotype frequencies. We have previously observed an equally high proportion of 10% in patients with schizophrenia (Foldager et al., 2012), but much larger samples would be needed to detect these differences as statistically significant deviations.

Inherited MBL deficiency presented by the XA/YO diplotype is more common (Thiel et al., 2006) and was seen in 9% of the controls (Table 1). A quite high number of patients with panic disorder (41%) were heterozygous for one of the mutations in exon 1 and almost half of these subjects (20%) carried the LXPA haplotype on the other strand, i.e. the XA/YO diplotype. In patients with bipolar disorder XA/YO was also relatively frequent (16%). These differences in proportions of XA/YO carriers as compared to the controls are significant for patients with panic disorder ($P=0.0048$) and just above the border of significance for patients with bipolar disorder ($P=0.062$). Note that LXPA and XA in the present study are equivalent and can be identified by the single variant X which was significantly associated with bipolar disorder, see Table 2.

Table 2
Trend tests (1 d.f. χ^2) for association of panic disorder and bipolar disorder with *MBL2* single locus and multilocus genetic markers by use of logistic regressions with an additive effect on a log scale of the minor allele (marked with bold type). Odds ratios (OR) measure the effect of each extra copy of the minor allele and OR between the two homozygote variants is therefore this value squared.

Markers	Panic disorder		Bipolar disorder	
	P value (adj.)[a]	OR (95% CI)	*P* value (adj.)[a]	OR (95% CI)
Single locus				
H/L (m1)	0.090 (0.41)	0.76 (0.55–1.04)	0.34 (0.90)	0.85 (0.62–1.18)
X/Y (m2)	0.16 (0.60)	1.31 (0.89–1.90)	**0.0075** (0.059)	1.65 (1.14–2.36)
P/**Q** (m3)	0.22 (0.67)	0.78 (0.51–1.15)	0.65 (0.98)	0.91 (0.61–1.34)
A/**D** (m4)	0.93 (0.94)	1.03 (0.55–1.84)	0.88 (0.95)	0.95 (0.50–1.72)
A/**B** (m5)	0.075 (0.41)	1.42 (0.96–2.07)	0.33 (0.90)	0.80 (0.50–1.24)
A/**C** (m6)	0.47 (0.84)	1.57 (0.42–4.95)	0.82 (0.98)	1.17 (0.26–4.00)
A/**O** (m7)[b]	0.067 –	1.38 (0.98–1.95)	0.39 –	0.85 (0.57–1.23)
Multilocus[c]				
HYPA	0.066 (0.41)	0.73 (0.51–1.02)	0.36 (0.89)	0.86 (0.61–1.19)
LYPA	0.79 (0.94)	1.09 (0.53–2.07)	0.56 (0.95)	0.80 (0.35–1.62)
LY**Q**A	0.14 (0.53)	0.73 (0.47–1.10)	0.59 (0.95)	0.90 (0.59–1.33)
YA[d]	**0.0074** –	0.66 (0.48–0.89)	0.14 –	0.79 (0.58–1.08)

[a] Permutation-based step-down max-statistics procedure accounting for nine simultaneous tests.
[b] The O allele of the A/O marker is any of the D, B and C variants of *MBL2* exon 1.
[c] LXPA, HYPD, LYPB and LYQC are identifiable with m2, m4, m5 and m6, respectively.
[d] XA and YO are identifiable with m2 and m7, respectively.

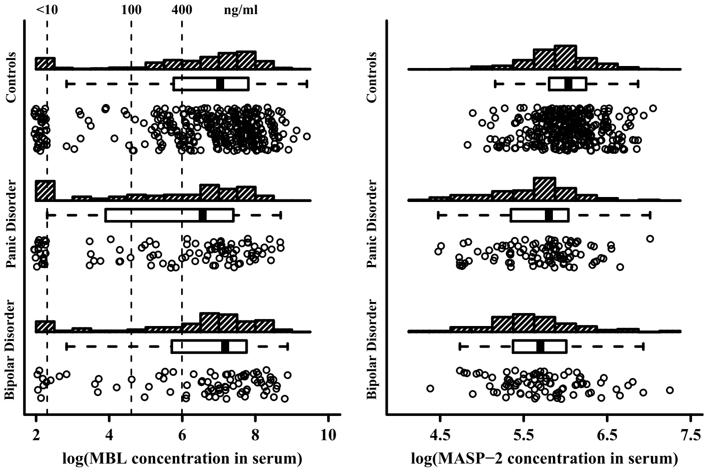


Fig. 1. Distribution of MBL and MASP-2 serum concentration. Concentration of MBL and MASP-2 in serum for 349 controls, 84 patients with bipolar disorder and 100 patients with panic disorder without a history of bipolar disorder. Before logarithmic transformation, concentrations were measured in ng protein per ml serum. The vertical lines in the left panel indicate: below MBL detection limit (< 10 ng/ml), low MBL level (< 100 ng/ml), intermediate MBL level (100–400 ng/ml) and normal MBL level (> 400 ng/ml). Histogram, box-plot and scatter plot of the observed concentrations are given for each protein and separately for each group of subjects.

3.4. Serum concentration

Fig. 1 presents the distribution of log-transformed MBL and MASP-2 serum concentrations. Significant *MBL2* single-marker and haplotype associations with MBL concentration were expected (Heitzeneder et al., 2012) and found (results not shown).

As shown in Table 3, significantly lower MBL serum concentration was found for patients with panic disorder as compared to controls, whereas no statistically significant difference was observed for bipolar disorder. These results remained when adjusting for *MBL2* haplotypes. Supplementary Tables S1 and S2 show median MBL concentrations estimated by these models and back-transformed using bootstrapping. To further investigate this, we compared quantiles of serum concentration from each case cohort with the controls, see Supplementary Table S3. The quantiles found in patients with panic disorder were in every case lower than in the controls although not always significantly. In patients with bipolar disorder the MBL concentrations were at about the same level as seen in the controls.

The serum concentrations of MASP-2 in patients with panic disorder as well as in patients with bipolar disorder were significantly lower than in controls, see Table 4. This was also evident from a supplementary investigation of the quantiles, see Supplementary Table S3.

From plots of MASP-2 serum concentration (not shown) we expected to find differences between cohorts and an effect of the *MASP2* SNP. The results from a forward inclusion procedure are shown in Table 4. We found the (logarithm of) MASP-2 serum

Table 3
Association of MBL serum concentration with panic and bipolar disorder, and after adjustment for the additive effects of carrying 0, 1 or 2 copies of each specific *MBL2* haplotype.

Models and parameters	Coefficient (95% CI)	Test statistic	*P* value
Panic disorder (PD)			
Intercept	6.366 (6.133 to 6.599)		
Phenotype (PD)	−0.826 (−1.322 to −0.330)	−3.26	**0.0011**
Haplotype model			
Intercept[a]	8.377 (8.186 to 8.569)		
Phenotype (PD)	−0.219 (−0.425 to −0.014)	−2.09	**0.037**
LYPA	−0.444 (−0.709 to −0.178)	−3.28	0.0010
LYQA	0.010 (−0.158 to 0.177)	0.11	0.91
LXPA	−1.336 (−1.501 to −1.170)	−15.8	2.4e−56
HYPD	−2.269 (−2.515 to −2.023)	−18.1	3.6e−73
LYPB	−3.495 (−3.693 to −3.297)	−34.6	2.9e−262
LYQC	−3.291 (−3.826 to −2.755)	−12.0	2.2e−33
Bipolar disorder (BD)			
Intercept	6.374 (6.153 to 6.595)		
Phenotype (BD)	0.074 (−0.427 to 0.575)	0.29	0.77
Haplotype model			
Intercept[a]	8.340 (8.149 to 8.530)		
Phenotype (BD)	0.115 (−0.097 to 0.328)	1.06	0.29
LYPA	−0.481 (−0.754 to −0.208)	−3.45	0.00056
LYQA	0.012 (−0.155 to 0.178)	0.14	0.89
LXPA	−1.300 (−1.465 to −1.136)	−15.5	3.3e−54
HYPD	−2.150 (−2.394 to −1.906)	−17.3	8.7e−67
LYPB	−3.447 (−3.649 to −3.244)	−33.3	1.1e−243
LYQC	−3.361 (−3.902 to −2.820)	−12.2	4.1e−34

[a] Controls with HYPA/HYPA multilocus genotype (i.e. two of the HYPA haplotype).

Table 4
Regression analyses of (log-transformed) MASP-2 concentration in serum for panic disorder (PD) versus controls and bipolar disorder (BD) versus controls. The final models are from a forward inclusion procedure using phenotype, carrier of the minor *MASP2* G-allele (A/G), *MBL2* single-markers, *MBL2* haplotypes and finally two-way interactions between parameters with significant main effects. Parameters are shown in order of inclusion.

Models and parameters	Coefficient (95% CI)	Test statistic[a]	*P* value
Panic disorder			
Intercept	6.020 (5.977 to 6.063)		
Phenotype (PD)	−0.327 (−0.418 to −0.236)	−7.07	6.2e−12
Final model			
Intercept	6.188 (6.122 to 6.253)		
MASP2	−0.485 (−0.618 to −0.353)	−7.19	2.7e−12
Phenotype (PD)	−0.323 (−0.408 to −0.238)	−7.47	4.2e−13
MBL2 YA	−0.107 (−0.154 to −0.061)	−4.56	6.7e−06
PD : *MASP2*[b]	−0.422 (−0.717 to −0.127)	−2.82	0.0051
Bipolar disorder			
Intercept	6.020 (5.977 to 6.063)		
Phenotype (BD)	−0.301 (−0.398 to −0.203)	−6.06	3.0e−9
Final model			
Intercept	6.204 (6.137 to 6.272)		
MASP2	−0.535 (−0.658 to −0.412)	−8.54	2.4e−16
Phenotype (BD)	−0.325 (−0.414 to −0.237)	−7.26	1.8e−12
MBL2 YA	−0.118 (−0.166 to −0.070)	−4.82	2.0e−06

[a] *P* value from Wald tests (H_0: parameter=0) evaluated in a t-distribution with degrees of freedom (d.f.) equal to the difference between the number of subjects (PD: 349+100=449 and BD: 349+84=433) and number of parameters in the model, e.g. 444 d.f. in final PD model.

[b] Here "V_1: V_2" represents the interaction effect between V_1 and V_2.

concentration to depend significantly on the *MASP2* SNP, disease phenotype and interestingly on the number of *MBL2* YA two-locus haplotypes. The interaction between phenotype and *MASP2* was also statistically significant for panic disorder but only on the border of significance (*P*=0.051) for bipolar disorder. Median MASP-2 concentrations estimated by these models are shown in Supplementary Tables S4 and S5.

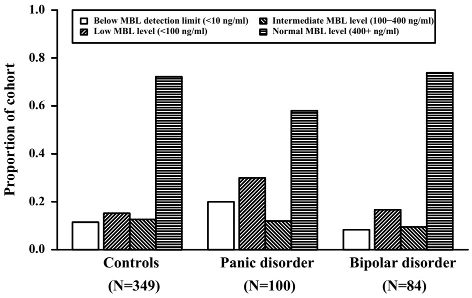


Fig. 2. Levels of MBL serum concentration. Bar plots of MBL serum levels for each of the three cohorts. The three hatched bars sums to one within each cohort corresponding to the categories: low MBL level (< 100 ng/ml), intermediate MBL level (100–400 ng/ml), and normal MBL level (> 400 ng/ml). The separated white bars indicate the proportion of measures that was below the detection limit (< 10 ng/ml) within each group. The proportion of measures below the detection limit was significantly higher (*P*=0.027) in patients with panic disorder as compared with controls (OR=1.9, CI: 1.1–3.5). Furthermore, the odds of having a low MBL level (< 100 ng/ml) was significantly increased (*P*=0.0008) for this group of patients with an odds ratio of 2.4 (CI: 1.4–4.0).

3.5. MBL serum deficiency

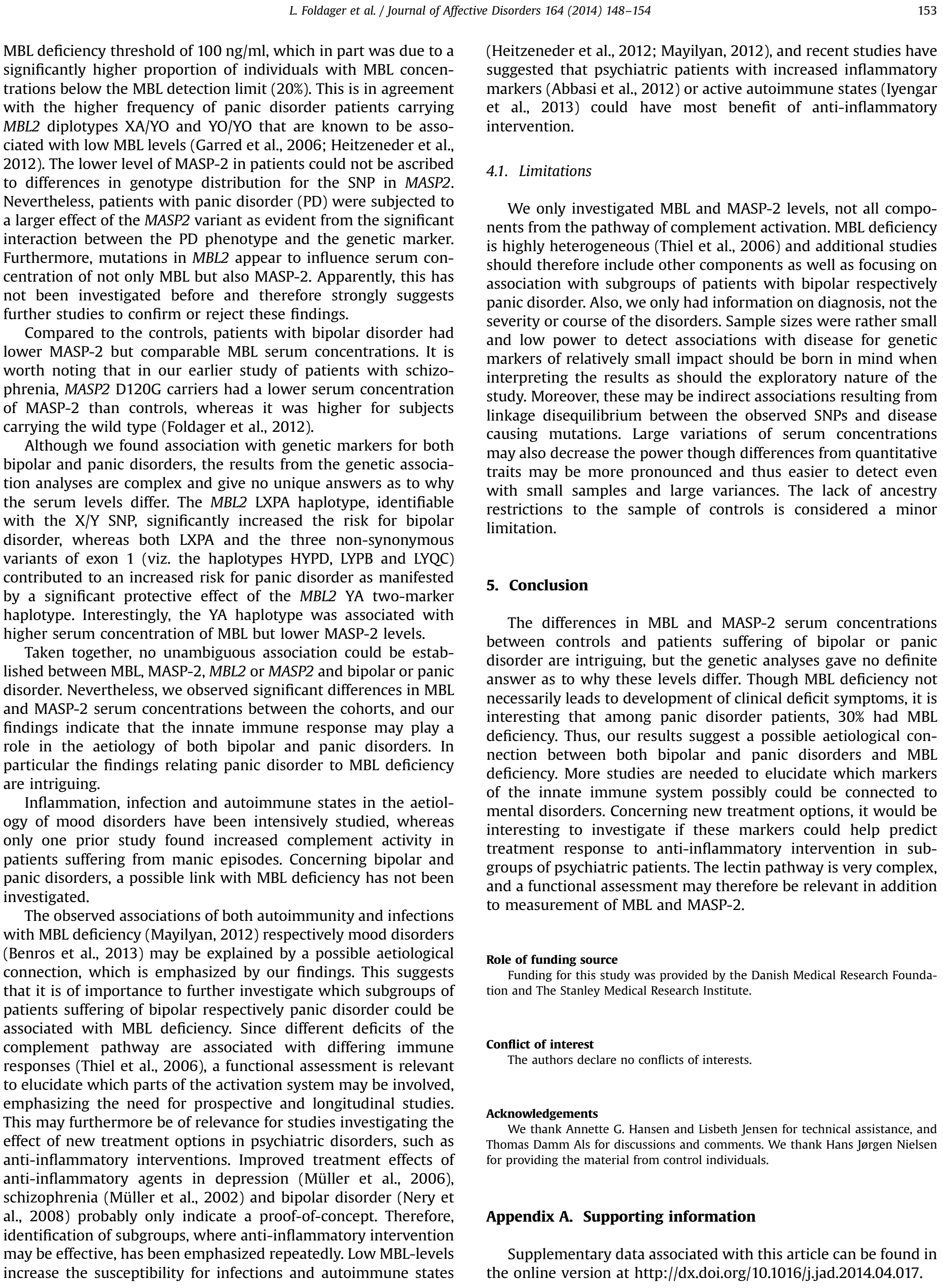
The hatched bars in Fig. 2 show the distributions according to the MBL serum concentration categories low (< 1 0 0), intermediate (100–400) and normal (> 400 ng/ml). Among patients with panic disorder 30% had a low MBL level while the corresponding proportions were 17% in patients with bipolar disorder and 15% in controls. This corresponds to an odds ratio of 2.4 (CI: 1.4–4.0; *P*=0.0008) when comparing panic disorder with controls.

Twenty of the patients with panic disorder (i.e. 20%) had undetectable MBL concentration, see the white bars in Fig. 2. This is a higher proportion than the 10–15% usually seen in general populations and significantly higher (*P*=0.027) than the 11% among controls that was observed in the present study: OR=1.9 (CI: 1.1–3.5). In patients with bipolar disorder we found a smaller proportion (8%) than usually seen but not significantly lower than in the controls.

4. Discussion

States of inflammation, infection and autoimmune diseases have been associated with the aetiology of psychiatric disorders (Benros et al., 2011, 2012, 2013; Krishnadas and Cavanagh, 2012; Leboyer et al., 2012) and different deficits within the pathway of complement activation (Heitzeneder et al., 2012; Mayilyan, 2012). Although several studies investigated the role of MBL in schizophrenia (Mayilyan et al., 2006; Spivak et al., 1993), only little research has been conducted to investigate implications of the innate immune system in mood disorders.

Thus, we investigated a possible connection between blood levels for MBL and MASP-2 and genetic markers for *MBL2* and *MASP2* with both bipolar and panic disorders. The most interesting finding was observed in patients with panic disorder, where the serum concentration of both MBL and MASP-2 was significantly lower than in controls. The proportion of controls with low (15%) and undetectable (11%) MBL-levels are in line with previous findings (Thiel et al., 2006). A significantly higher proportion of patients with panic disorder (30%) had levels below the defined



MBL deficiency threshold of 100 ng/ml, which in part was due to a significantly higher proportion of individuals with MBL concentrations below the MBL detection limit (20%). This is in agreement with the higher frequency of panic disorder patients carrying *MBL2* diplotypes XA/YO and YO/YO that are known to be associated with low MBL levels (Garred et al., 2006; Heitzeneder et al., 2012). The lower level of MASP-2 in patients could not be ascribed to differences in genotype distribution for the SNP in *MASP2*. Nevertheless, patients with panic disorder (PD) were subjected to a larger effect of the *MASP2* variant as evident from the significant interaction between the PD phenotype and the genetic marker. Furthermore, mutations in *MBL2* appear to influence serum concentration of not only MBL but also MASP-2. Apparently, this has not been investigated before and therefore strongly suggests further studies to confirm or reject these findings.

Compared to the controls, patients with bipolar disorder had lower MASP-2 but comparable MBL serum concentrations. It is worth noting that in our earlier study of patients with schizophrenia, *MASP2* D120G carriers had a lower serum concentration of MASP-2 than controls, whereas it was higher for subjects carrying the wild type (Foldager et al., 2012).

Although we found association with genetic markers for both bipolar and panic disorders, the results from the genetic association analyses are complex and give no unique answers as to why the serum levels differ. The *MBL2* LXPA haplotype, identifiable with the X/Y SNP, significantly increased the risk for bipolar disorder, whereas both LXPA and the three non-synonymous variants of exon 1 (viz. the haplotypes HYPD, LYPB and LYQC) contributed to an increased risk for panic disorder as manifested by a significant protective effect of the *MBL2* YA two-marker haplotype. Interestingly, the YA haplotype was associated with higher serum concentration of MBL but lower MASP-2 levels.

Taken together, no unambiguous association could be established between MBL, MASP-2, *MBL2* or *MASP2* and bipolar or panic disorder. Nevertheless, we observed significant differences in MBL and MASP-2 serum concentrations between the cohorts, and our findings indicate that the innate immune response may play a role in the aetiology of both bipolar and panic disorders. In particular the findings relating panic disorder to MBL deficiency are intriguing.

Inflammation, infection and autoimmune states in the aetiology of mood disorders have been intensively studied, whereas only one prior study found increased complement activity in patients suffering from manic episodes. Concerning bipolar and panic disorders, a possible link with MBL deficiency has not been investigated.

The observed associations of both autoimmunity and infections with MBL deficiency (Mayilyan, 2012) respectively mood disorders (Benros et al., 2013) may be explained by a possible aetiological connection, which is emphasized by our findings. This suggests that it is of importance to further investigate which subgroups of patients suffering of bipolar respectively panic disorder could be associated with MBL deficiency. Since different deficits of the complement pathway are associated with differing immune responses (Thiel et al., 2006), a functional assessment is relevant to elucidate which parts of the activation system may be involved, emphasizing the need for prospective and longitudinal studies. This may furthermore be of relevance for studies investigating the effect of new treatment options in psychiatric disorders, such as anti-inflammatory interventions. Improved treatment effects of anti-inflammatory agents in depression (Müller et al., 2006), schizophrenia (Müller et al., 2002) and bipolar disorder (Nery et al., 2008) probably only indicate a proof-of-concept. Therefore, identification of subgroups, where anti-inflammatory intervention may be effective, has been emphasized repeatedly. Low MBL-levels increase the susceptibility for infections and autoimmune states (Heitzeneder et al., 2012; Mayilyan, 2012), and recent studies have suggested that psychiatric patients with increased inflammatory markers (Abbasi et al., 2012) or active autoimmune states (Iyengar et al., 2013) could have most benefit of anti-inflammatory intervention.

4.1. Limitations

We only investigated MBL and MASP-2 levels, not all components from the pathway of complement activation. MBL deficiency is highly heterogeneous (Thiel et al., 2006) and additional studies should therefore include other components as well as focusing on association with subgroups of patients with bipolar respectively panic disorder. Also, we only had information on diagnosis, not the severity or course of the disorders. Sample sizes were rather small and low power to detect associations with disease for genetic markers of relatively small impact should be born in mind when interpreting the results as should the exploratory nature of the study. Moreover, these may be indirect associations resulting from linkage disequilibrium between the observed SNPs and disease causing mutations. Large variations of serum concentrations may also decrease the power though differences from quantitative traits may be more pronounced and thus easier to detect even with small samples and large variances. The lack of ancestry restrictions to the sample of controls is considered a minor limitation.

5. Conclusion

The differences in MBL and MASP-2 serum concentrations between controls and patients suffering of bipolar or panic disorder are intriguing, but the genetic analyses gave no definite answer as to why these levels differ. Though MBL deficiency not necessarily leads to development of clinical deficit symptoms, it is interesting that among panic disorder patients, 30% had MBL deficiency. Thus, our results suggest a possible aetiological connection between both bipolar and panic disorders and MBL deficiency. More studies are needed to elucidate which markers of the innate immune system possibly could be connected to mental disorders. Concerning new treatment options, it would be interesting to investigate if these markers could help predict treatment response to anti-inflammatory intervention in subgroups of psychiatric patients. The lectin pathway is very complex, and a functional assessment may therefore be relevant in addition to measurement of MBL and MASP-2.

Role of funding source

Funding for this study was provided by the Danish Medical Research Foundation and The Stanley Medical Research Institute.

Conflict of interest

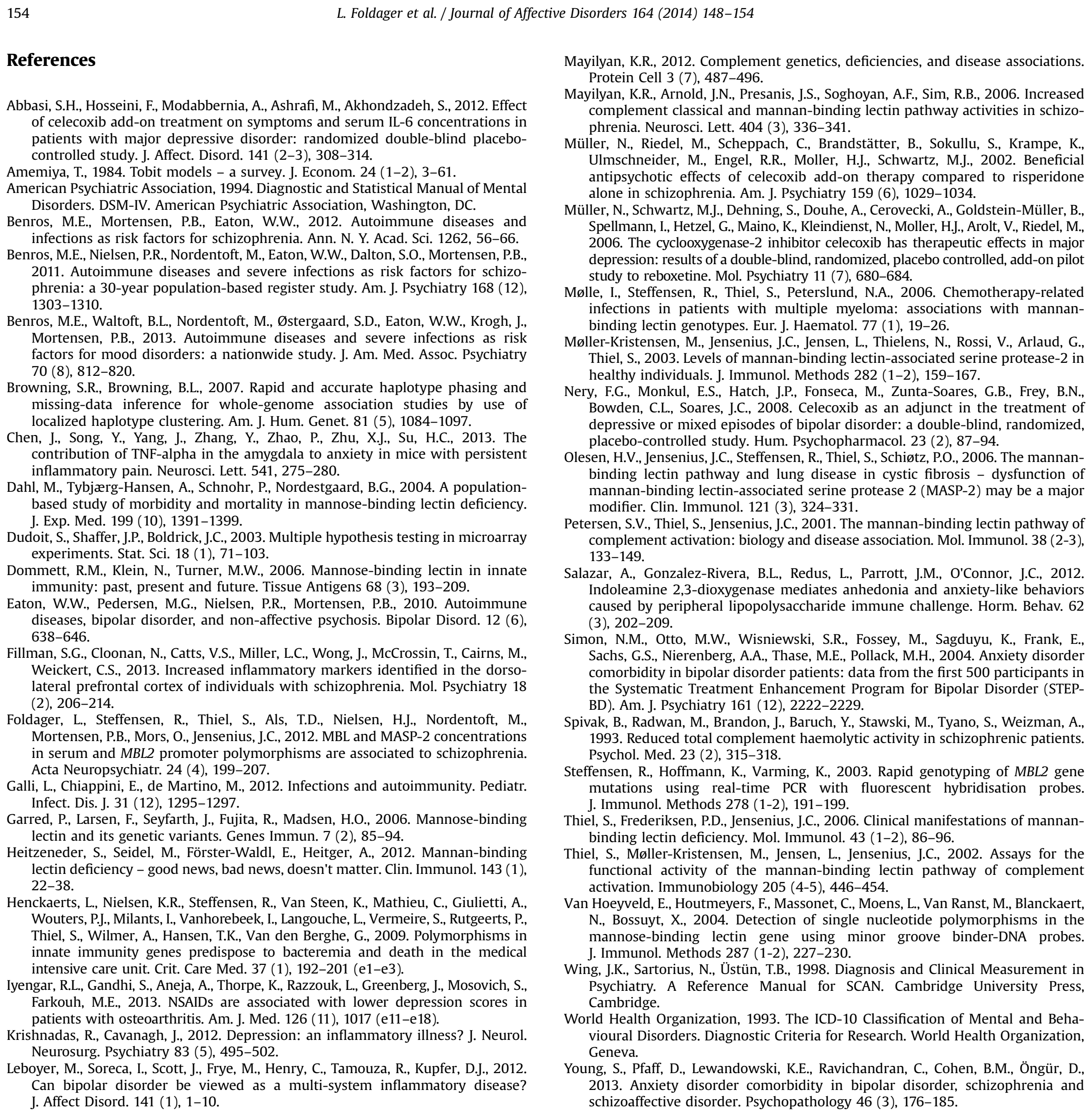
The authors declare no conflicts of interests.

Acknowledgements

We thank Annette G. Hansen and Lisbeth Jensen for technical assistance, and Thomas Damm Als for discussions and comments. We thank Hans Jørgen Nielsen for providing the material from control individuals.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jad.2014.04.017.



References

Abbasi, S.H., Hosseini, F., Modabbernia, A., Ashrafi, M., Akhondzadeh, S., 2012. Effect of celecoxib add-on treatment on symptoms and serum IL-6 concentrations in patients with major depressive disorder: randomized double-blind placebo-controlled study. J. Affect. Disord. 141 (2–3), 308–314.

Amemiya, T., 1984. Tobit models – a survey. J. Econom. 24 (1–2), 3–61.

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders. DSM-IV. American Psychiatric Association, Washington, DC.

Benros, M.E., Mortensen, P.B., Eaton, W.W., 2012. Autoimmune diseases and infections as risk factors for schizophrenia. Ann. N. Y. Acad. Sci. 1262, 56–66.

Benros, M.E., Nielsen, P.R., Nordentoft, M., Eaton, W.W., Dalton, S.O., Mortensen, P.B., 2011. Autoimmune diseases and severe infections as risk factors for schizophrenia: a 30-year population-based register study. Am. J. Psychiatry 168 (12), 1303–1310.

Benros, M.E., Waltoft, B.L., Nordentoft, M., Østergaard, S.D., Eaton, W.W., Krogh, J., Mortensen, P.B., 2013. Autoimmune diseases and severe infections as risk factors for mood disorders: a nationwide study. J. Am. Med. Assoc. Psychiatry 70 (8), 812–820.

Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81 (5), 1084–1097.

Chen, J., Song, Y., Yang, J., Zhang, Y., Zhao, P., Zhu, X.J., Su, H.C., 2013. The contribution of TNF-alpha in the amygdala to anxiety in mice with persistent inflammatory pain. Neurosci. Lett. 541, 275–280.

Dahl, M., Tybjærg-Hansen, A., Schnohr, P., Nordestgaard, B.G., 2004. A population-based study of morbidity and mortality in mannose-binding lectin deficiency. J. Exp. Med. 199 (10), 1391–1399.

Dudoit, S., Shaffer, J.P., Boldrick, J.C., 2003. Multiple hypothesis testing in microarray experiments. Stat. Sci. 18 (1), 71–103.

Dommett, R.M., Klein, N., Turner, M.W., 2006. Mannose-binding lectin in innate immunity: past, present and future. Tissue Antigens 68 (3), 193–209.

Eaton, W.W., Pedersen, M.G., Nielsen, P.R., Mortensen, P.B., 2010. Autoimmune diseases, bipolar disorder, and non-affective psychosis. Bipolar Disord. 12 (6), 638–646.

Fillman, S.G., Cloonan, N., Catts, V.S., Miller, L.C., Wong, J., McCrossin, T., Cairns, M., Weickert, C.S., 2013. Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. Mol. Psychiatry 18 (2), 206–214.

Foldager, L., Steffensen, R., Thiel, S., Als, T.D., Nielsen, H.J., Nordentoft, M., Mortensen, P.B., Mors, O., Jensenius, J.C., 2012. MBL and MASP-2 concentrations in serum and *MBL2* promoter polymorphisms are associated to schizophrenia. Acta Neuropsychiatr. 24 (4), 199–207.

Galli, L., Chiappini, E., de Martino, M., 2012. Infections and autoimmunity. Pediatr. Infect. Dis. J. 31 (12), 1295–1297.

Garred, P., Larsen, F., Seyfarth, J., Fujita, R., Madsen, H.O., 2006. Mannose-binding lectin and its genetic variants. Genes Immun. 7 (2), 85–94.

Heitzeneder, S., Seidel, M., Förster-Waldl, E., Heitger, A., 2012. Mannan-binding lectin deficiency – good news, bad news, doesn't matter. Clin. Immunol. 143 (1), 22–38.

Henckaerts, L., Nielsen, K.R., Steffensen, R., Van Steen, K., Mathieu, C., Giulietti, A., Wouters, P.J., Milants, I., Vanhorebeek, I., Langouche, L., Vermeire, S., Rutgeerts, P., Thiel, S., Wilmer, A., Hansen, T.K., Van den Berghe, G., 2009. Polymorphisms in innate immunity genes predispose to bacteremia and death in the medical intensive care unit. Crit. Care Med. 37 (1), 192–201 (e1–e3).

Iyengar, R.L., Gandhi, S., Aneja, A., Thorpe, K., Razzouk, L., Greenberg, J., Mosovich, S., Farkouh, M.E., 2013. NSAIDs are associated with lower depression scores in patients with osteoarthritis. Am. J. Med. 126 (11), 1017 (e11–e18).

Krishnadas, R., Cavanagh, J., 2012. Depression: an inflammatory illness? J. Neurol. Neurosurg. Psychiatry 83 (5), 495–502.

Leboyer, M., Soreca, I., Scott, J., Frye, M., Henry, C., Tamouza, R., Kupfer, D.J., 2012. Can bipolar disorder be viewed as a multi-system inflammatory disease? J. Affect Disord. 141 (1), 1–10.

Mayilyan, K.R., 2012. Complement genetics, deficiencies, and disease associations. Protein Cell 3 (7), 487–496.

Mayilyan, K.R., Arnold, J.N., Presanis, J.S., Soghoyan, A.F., Sim, R.B., 2006. Increased complement classical and mannan-binding lectin pathway activities in schizophrenia. Neurosci. Lett. 404 (3), 336–341.

Müller, N., Riedel, M., Scheppach, C., Brandstätter, B., Sokullu, S., Krampe, K., Ulmschneider, M., Engel, R.R., Moller, H.J., Schwartz, M.J., 2002. Beneficial antipsychotic effects of celecoxib add-on therapy compared to risperidone alone in schizophrenia. Am. J. Psychiatry 159 (6), 1029–1034.

Müller, N., Schwartz, M.J., Dehning, S., Douhe, A., Cerovecki, A., Goldstein-Müller, B., Spellmann, I., Hetzel, G., Maino, K., Kleindienst, N., Moller, H.J., Arolt, V., Riedel, M., 2006. The cyclooxygenase-2 inhibitor celecoxib has therapeutic effects in major depression: results of a double-blind, randomized, placebo controlled, add-on pilot study to reboxetine. Mol. Psychiatry 11 (7), 680–684.

Mølle, I., Steffensen, R., Thiel, S., Peterslund, N.A., 2006. Chemotherapy-related infections in patients with multiple myeloma: associations with mannan-binding lectin genotypes. Eur. J. Haematol. 77 (1), 19–26.

Møller-Kristensen, M., Jensenius, J.C., Jensen, L., Thielens, N., Rossi, V., Arlaud, G., Thiel, S., 2003. Levels of mannan-binding lectin-associated serine protease-2 in healthy individuals. J. Immunol. Methods 282 (1–2), 159–167.

Nery, F.G., Monkul, E.S., Hatch, J.P., Fonseca, M., Zunta-Soares, G.B., Frey, B.N., Bowden, C.L., Soares, J.C., 2008. Celecoxib as an adjunct in the treatment of depressive or mixed episodes of bipolar disorder: a double-blind, randomized, placebo-controlled study. Hum. Psychopharmacol. 23 (2), 87–94.

Olesen, H.V., Jensenius, J.C., Steffensen, R., Thiel, S., Schiøtz, P.O., 2006. The mannan-binding lectin pathway and lung disease in cystic fibrosis – dysfunction of mannan-binding lectin-associated serine protease 2 (MASP-2) may be a major modifier. Clin. Immunol. 121 (3), 324–331.

Petersen, S.V., Thiel, S., Jensenius, J.C., 2001. The mannan-binding lectin pathway of complement activation: biology and disease association. Mol. Immunol. 38 (2-3), 133–149.

Salazar, A., Gonzalez-Rivera, B.L., Redus, L., Parrott, J.M., O'Connor, J.C., 2012. Indoleamine 2,3-dioxygenase mediates anhedonia and anxiety-like behaviors caused by peripheral lipopolysaccharide immune challenge. Horm. Behav. 62 (3), 202–209.

Simon, N.M., Otto, M.W., Wisniewski, S.R., Fossey, M., Sagduyu, K., Frank, E., Sachs, G.S., Nierenberg, A.A., Thase, M.E., Pollack, M.H., 2004. Anxiety disorder comorbidity in bipolar disorder patients: data from the first 500 participants in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). Am. J. Psychiatry 161 (12), 2222–2229.

Spivak, B., Radwan, M., Brandon, J., Baruch, Y., Stawski, M., Tyano, S., Weizman, A., 1993. Reduced total complement haemolytic activity in schizophrenic patients. Psychol. Med. 23 (2), 315–318.

Steffensen, R., Hoffmann, K., Varming, K., 2003. Rapid genotyping of *MBL2* gene mutations using real-time PCR with fluorescent hybridisation probes. J. Immunol. Methods 278 (1-2), 191–199.

Thiel, S., Frederiksen, P.D., Jensenius, J.C., 2006. Clinical manifestations of mannan-binding lectin deficiency. Mol. Immunol. 43 (1–2), 86–96.

Thiel, S., Møller-Kristensen, M., Jensen, L., Jensenius, J.C., 2002. Assays for the functional activity of the mannan-binding lectin pathway of complement activation. Immunobiology 205 (4-5), 446–454.

Van Hoeyveld, E., Houtmeyers, F., Massonet, C., Moens, L., Van Ranst, M., Blanckaert, N., Bossuyt, X., 2004. Detection of single nucleotide polymorphisms in the mannose-binding lectin gene using minor groove binder-DNA probes. J. Immunol. Methods 287 (1-2), 227–230.

Wing, J.K., Sartorius, N., Üstün, T.B., 1998. Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN. Cambridge University Press, Cambridge.

World Health Organization, 1993. The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research. World Health Organization, Geneva.

Young, S., Pfaff, D., Lewandowski, K.E., Ravichandran, C., Cohen, B.M., Öngür, D., 2013. Anxiety disorder comorbidity in bipolar disorder, schizophrenia and schizoaffective disorder. Psychopathology 46 (3), 176–185.

6.2.1 Paper 2 - supplementary info

L. Foldager, O. Köhler, R. Steffensen, S. Thiel, A.S. Kristensen, J.C. Jensenius, O. Mors. *Bipolar and panic disorders may be associated with hereditary defects in the innate immune system.*

Supplementary Information

Supplementary table legends

Supplementary Table S1. Estimated median MBL concentration in serum for patients with panic disorder.

Median MBL concentration in serum (ng/ml) estimated from Tobit regressions (on log-transformed data) with patients and controls specific means and an additive effect of haplotypes (Table 3). The 95% confidence intervals in the parentheses were estimated by use of the normal approximation on results from ordinary bootstrapping with 10,000 replicates (Davison and Hinkley, 1997). Results from using only the two markers X/Y and A/O (see Table 1) are given before the corresponding four-marker multilocus genotype groups (e.g. YA/YA corresponds to YA=2, XA=YO=0). The results in the first row (Any) are from the model with only patient/control status as a factor.

Supplementary Table S2. Estimated median MBL concentration in serum for patients with bipolar disorder.

Median MBL concentration in serum (ng/ml) estimated from Tobit regressions (on log-transformed data) with patients and controls specific means and an additive effect of haplotypes (Table 3). The 95% confidence intervals in the parentheses were estimated by use of the normal approximation on results from ordinary bootstrapping with 10,000 replicates (Davison and Hinkley, 1997). Results from using only the two markers X/Y and A/O (see Table 1) are given before the corresponding four-marker multilocus genotype groups (e.g. YA/YA corresponds to YA=2, XA=YO=0). The results in the first row (Any) are from the model with only patient/control status as a factor.

Supplementary Table S3. Quantiles of MBL and MASP-2 concentration in serum. Comparisons with the controls were carried out using an extended version of the usual median test (Conover, 1999). P values were based on Monte Carlo simulations (Hope, 1968) using 100,000,000 (i.e. 1e8) replicates. MBL concentrations below the detection limit were set equal to this 10 ng/ml limit.

Supplementary Table S4. Estimated median MASP-2 concentration in serum for patients with panic disorder.

Median MASP-2 concentration in serum (ng/ml) estimated from regressions analysis (on log-transformed data) with patients, controls and *MASP2* genotype specific means (interaction effect) and a linear effect of the *MBL2* YA haplotype (Table 4, panic disorder final model). The 95% confidence intervals (in the parentheses) were estimated by use of the normal approximation on results from ordinary bootstrapping with 10,000 replicates (Davison and Hinkley, 1997). Results obtained without inclusion of the *MBL2* haplotype are given before the corresponding combination with *MBL2* multilocus genotype (*MASP2* A/A and *MASP2* A/G rows). The results in the first row (Any) are from the model which only includes the patient/control factor (Table 4, panic disorder), i.e. any genotype combination.

Supplementary Table S5. Estimated median MASP-2 concentration in serum for patients with bipolar disorder.

Median MASP-2 concentration in serum (ng/ml) estimated from regressions analysis (on log-transformed data) with patients and controls specific means and linear effects of *MASP2* G allele and *MBL2* YA haplotype (Table 4, bipolar disorder final model). The 95% confidence intervals (in the parentheses) were estimated by use of the normal approximation on results from ordinary bootstrapping with 10,000 replicates (Davison and Hinkley, 1997). Results obtained without inclusion of the *MBL2* haplotype are given before the corresponding combination with *MBL2* multilocus genotype (*MASP2* A/A and *MASP2* A/G rows). The results in the first row (Any) are from the model which only includes the patient/control factor (Table 4, bipolar disorder), i.e. any genotype combination.

Supplementary references

Conover, W.J., 1999. Practical nonparametric statistics. Wiley, New York, NY.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap methods and their application. Cambridge University Press, Cambridge.

Hope, A.C.A., 1968. A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. Series B Stat. Methodol. 30 (3), 582-598.

Foldager et al.

Supplementary Table S1. Estimated median MBL concentration in serum for patients with panic disorder.

MBL2 genotype	Controls (N = 349)	Panic disorder (N = 100)
Any	582 (442 - 717)	255 (113 - 380)
YA/YA	4238 (3684 - 4769)	3274 (2451 - 4043)
HYPA/HYPA	4346 (3636 - 5027)	3491 (2550 - 4375)
HYPA/LYPA	2789 (1810 - 3684)	2240 (1409 - 3000)
HYPA/LYQA	4388 (3855 - 4906)	3524 (2651 - 4352)
LYPA/LYPA	1789 (469 - 2912)	1437 (405 - 2319)
LYQA/LYPA	2816 (1835 - 3715)	2261 (1426 - 3027)
LYQA/LYQA	4430 (3497 - 5321)	3558 (2462 - 4585)
YA/XA	1104 (967 - 1238)	853 (654 - 1041)
HYPA/LXPA	1143 (1000 - 1279)	918 (708 - 1116)
LYPA/LXPA	733 (469 - 974)	589 (373 - 786)
LYQA/LXPA	1154 (981 - 1319)	927 (698 - 1142)
XA/XA	289 (210 - 361)	222 (150 - 290)
LXPA/LXPA	301 (227 - 369)	241 (169 - 308)
YA/YO	190 (157 - 221)	147 (108 - 183)
HYPA/LYPB	132 (103 - 160)	106 (76 - 134)
LYPA/LYPB	85 (50 - 116)	68 (40 - 93)
LYQA/LYPB	133 (104 - 161)	107 (77 - 136)
HYPA/LYQC	162 (52 - 257)	130 (38 - 210)
LYQA/LYQC	163 (52 - 261)	131 (38 - 213)
HYPA/HYPD	449 (306 - 580)	361 (218 - 490)
LYPA /HYPD	288 (155 - 406)	232 (119 - 331)
LYQA /HYPD	454 (312 - 583)	364 (222 - 493)
XA/YO	50 (38 - 61)	38 (27 - 49)
LXPA/HYPD	118 (79 - 154)	95 (57 - 129)
LXPA/LYPB	35 (26 - 43)	28 (19 - 36)
LXPA/LYQC	43 (12 - 69)	34 (9 - 56)
YO/YO	9 (6 - 11)	7 (4 - 9)
HYPD/HYPD	46 (15 - 73)	37 (10 - 60)
HYPD/LYPB	14 (8 - 18)	11 (6 - 15)
HYPD/LYQC	17 (3 - 28)	13 (2 - 23)
LYPB/LYPB	4 (2 - 6)	3 (2 - 5)
LYPB/LYQC	5 (1 - 8)	4 (1 - 6)

Supplementary Table S2. Estimated median MBL concentration in serum for patients with bipolar disorder.

MBL2 genotype	Controls (N = 349)	Bipolar disorder (N = 84)
Any	586 (448 - 718)	631 (322 - 907)
YA/YA	4030 (3497 - 4541)	4716 (3521 - 5847)
HYPA/HYPA	4187 (3519 - 4840)	4699 (3507 - 5834)
HYPA/LYPA	2589 (1605 - 3477)	2906 (1667 - 4016)
HYPA/LYQA	4237 (3726 - 4746)	4756 (3680 - 5789)
LYPA/LYPA	1601 (295 - 2686)	1797 (285 - 3050)
LYQA/LYPA	2620 (1638 - 3509)	2941 (1707 - 4049)
LYQA/LYQA	4288 (3423 - 5125)	4813 (3476 - 6079)
YA/XA	1092 (952 - 1228)	1278 (951 - 1589)
HYPA/LXPA	1141 (996 - 1280)	1280 (971 - 1571)
LYPA/LXPA	705 (428 - 952)	791 (441 - 1102)
LYQA/LXPA	1154 (985 - 1317)	1296 (972 - 1600)
XA/XA	296 (212 - 375)	346 (218 - 466)
LXPA/LXPA	311 (230 - 385)	349 (229 - 458)
YA/YO	196 (160 - 230)	230 (164 - 291)
HYPA/LYPB	133 (102 - 162)	150 (106 - 190)
LYPA/LYPB	82 (46 - 115)	93 (48 - 131)
LYQA/LYPB	135 (104 - 164)	151 (108 - 192)
HYPA/LYQC	145 (24 - 248)	163 (20 - 284)
LYQA/LYQC	147 (25 - 251)	165 (21 - 287)
HYPA/HYPD	488 (334 - 626)	547 (348 - 725)
LYPA /HYPD	302 (154 - 429)	338 (163 - 489)
LYQA /HYPD	493 (338 - 633)	554 (354 - 733)
XA/YO	53 (40 - 66)	62 (41 - 82)
LXPA/HYPD	133 (87 - 173)	149 (90 - 201)
LXPA/LYPB	36 (26 - 45)	41 (27 - 53)
LXPA/LYQC	40 (5 - 68)	44 (4 - 79)
YO/YO	10 (6 - 13)	11 (6 - 16)
HYPD/HYPD	57 (18 - 89)	64 (19 - 101)
HYPD/LYPB	16 (9 - 21)	17 (10 - 24)
HYPD/LYQC	17 (1 - 30)	19 (0 - 34)
LYPB/LYPB	4 (2 - 6)	5 (2 - 7)
LYPB/LYQC	5 (1 - 8)	5 (0 - 9)

Foldager et al.

Supplementary Table S3. Quantiles of MBL and MASP-2 concentration in serum.

	Probability	Controls quantile	Panic disorder quantile (χ^2, P value)	Bipolar disorder quantile (χ^2, P value)
MBL	0.1	10	10 (4.90, 0.031)	15 (0.68, 0.45)
	0.25	319	54 (11.2, 0.0011)	309 (0.057, 0.89)
	0.5 (median)	1133	704 (4.59, 0.041)	1307 (0.26, 0.63)
	0.75	2460	1629 (3.31, 0.088)	2333 (0.071, 0.89)
	0.9	3809	2921 (1.30, 0.27)	4285 (0.98, 0.42)
MASP-2	0.1	260	127 (39.3, 3.0e-8)	165 (32.3, 3.3e-7)
	0.25	332	210 (24.2, 2.1e-6)	218 (48.4, 1.0e-8[a])
	0.5 (median)	417	331 (25.9, 5.0e-7)	299 (25.8, 4.4e-7)
	0.75	517	418 (13.0, 3.4e-4)	403 (6.32, 0.016)
	0.9	664	549 (5.17, 0.023)	596 (1.04, 0.33)

a) P=1.0e-8 from 1e8 replicates means that none of the replicates were more extreme than the observed so in reality P may well be smaller.

Supplementary Table S4. Estimated median MASP-2 concentration in serum for patients with panic disorder.

MBL2 genotype	Controls (N = 349)	Panic disorder (N = 100)
Any	412 (396 - 428)	297 (267 - 326)
***MASP2* A/A**	431 (415 - 447)	319 (289 - 348)
YA/YA	397 (376 - 418)	278 (251 - 304)
YA/YO	440 (424 - 457)	308 (281 - 335)
YO/YO	488 (458 - 519)	342 (307 - 376)
***MASP2* A/G**	260 (223 - 296)	131 (96 - 164)
YA/YA	224 (194 - 253)	157 (132 - 180)
YA/YO	249 (216 - 280)	174 (147 - 200)
YO/YO	276 (236 - 314)	193 (161 - 224)

Supplementary Table S5. Estimated median MASP-2 concentration in serum for patients with bipolar disorder.

MBL2 genotype	Controls (N = 349)	Bipolar disorder (N = 86)
Any	412 (396 - 428)	305 (270 - 338)
***MASP2* A/A**	433 (417 - 450)	317 (283 - 350)
YA/YA	391 (369 - 412)	282 (251 - 313503)
YA/YO	440 (424 - 456)	318 (284 - 350)
YO/YO	595 (463 - 527)	357 (315 - 399)
***MASP2* A/G**	249 (218 - 280)	182 (155 - 209)
YA/YA	229 (200 - 257)	165 (140 - 189)
YA/YO	258 (226 - 289)	186 (158 - 213)
YO/YO	290 (250 - 329)	209 (175 - 243)

6.3 Paper 3[32]

Comparison of methods for genome-wide gene-environment interaction analysis

Leslie Foldager[1,2,3,4]*, Thomas Damm Als[3,4,5] and Jakob Grove[2,3,4,5]

[1] Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Risskov, Denmark
[2] Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[3] *i*PSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus and Copenhagen, Denmark
[4] *i*SEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark
[5] Department of Biomedicine, Aarhus University, Aarhus, Denmark

Aarhus University, 31 March 2014

Abstract

Background Gene-environment ($G \times E$) interaction may be an important source of complexity in the aetiology of psychiatric disorders but so far only few interactions have been reported, possibly due to the need of large samples in concert with information on the environmental exposure. The Danish *i*PSYCH study will include such information. On the other hand, the number of methods and software available for $G \times E$ analysis is overwhelming and none appears to be superior. To guide decisions in *i*PSYCH we therefore set up a simulation study to compare some of the most promising or frequently used approaches. The manuscript in its present form describes the construction of the data sets that are to form the test bed for the comparisons of methods that are to follow.

Methods Case-control samples with individual-based genetic data were simulated by use of simuPOP scripts allowing to model penetrances under restrictions specified by biological and epidemiological parameters and with interactions between up to two disease predisposing genetic markers and one predisposing environmental factor. The methods for analysing $G \times E$ interaction investigated in the present study are traditional two-step logistic regression, *logicFS* and *MB-MDR*.

Results From an initial population of 993 unrelated HapMap3 individuals and a selected set of SNPs, a *base population* of 50,000 individuals was generated by linear expansion of the initial population for 500 non-overlapping generations. By repeatedly generating one offspring from the base population by random mating and imposing affection status under certain scenarios, case-control samples were drawn using rejection sampling. The transmission of genotypes in this process was subject to the same evolutionary parameters as used for expansion of the initial population. One hundred samples of 5,000 affected and 5,000 unaffected individuals were generated for 16 scenarios.

Discussion The current status of the study is that case-control samples have been generated and analyses using the three $G \times E$ methods initiated. When applying machine learning methods in practice several different methods, algorithms and/or sets of parameters are often used. So far we have just used default settings if possible or reasonable but a more thorough investigation should be made to choose optimal algorithmic parameters.

Funding and disclaimer This study was funded by the Lundbeck Foundation, Denmark. The Lundbeck Foundation had no involvement in any aspect of the study.

Keywords Gene-environment interaction; machine learning; MB-MDR; logic regression; logicFS; two-step methods, simulation study

*Corresponding author: leslie@birc.au.dk

[32] Manuscript in preparation.

1 Introduction

Gene-environment interaction (G×E) may be an important source of complexity to the aetiology of psychiatric disorders. Nevertheless, only few findings have been reported, possibly due to at least two complicating factors: the need of large samples in concert with information on the environmental exposure. The comprehensive Danish study *i*PSYCH[1] includes such information drawing partly on the Danish registers and partly on the ability to extract information from neonatal dried blood spots about exposures to the fetus. An example of successful application of data from the same source was recently given by Borglum et al. (2013) showing a statistically significant $G \times E$ interaction with maternal infection by cytomegalovirus and thereby suggesting new susceptibility loci for schizophrenia.

Choosing methods and software to be used for disentangling these possible G×E interactions can be difficult as the abundance of different methods and software is overwhelming and with no obvious gold standard. In Borglum et al. (2013) the two-step method by Murcray et al. (2009) was applied and multi-step approaches like that represents one option. However these are often just searching for fairly simply interaction—typically just two-way interactions between one or a few environmental factors or covariates and a collection of single nucleotide polymorphisms (SNPs) selected from a genome-wide association study (GWAS). To guide decision of which G×E methods to use, we decided to set up a simulation study to compare some of the most promising or frequently used methods, from multi-step regression analyses to machine learning. The intention being to characterize performance of a number of G×E methods in a wide range of standardized scenarios to facilitate informed choices in future and ongoing projects such as *i*PSYCH[1]. We intend to consider a range of scenarios by varying minor allele frequencies (MAFs), sample sizes, models and effect sizes with the intention to compare methods of the following kind: two-step analysis, multifactor dimensionality reduction (MDR) (Ritchie et al., 2001), logic regression (Ruczinski et al., 2003), random forests (Breiman, 2001), artificial neural networks (ANN or just NN), genetic programmed neural networks (GPNN). The idea of artificial neural networks dates back the 1940's but for a review on the use of NN's in genetic epidemiology also covering GPNN's we refer to Motsinger-Reif et al. (2008).

We present here the first steps of a larger simulation study. The aims of this part of the study were: 1) Decide on methods/software for simulation of individual-based genetic data and generation of case-control samples with phenotype assignment (affection status determination) subject to penetrances dependent on G×E and G×G interactions; 2) Simulate samples for a small set of scenarios and check that the data generated complies with assumed models and parameters; 3) Analyse the generated data with a set of G×E methods to a) check technical and scientific performance of a few of our first-line choices of G×E methods, b) determine how to meaningfully compare methods that are very different in nature, and c) devise the range of scenarios to be used for the larger study. The present manuscript focuses on the first two aims of the study. To meet intended requirements for the simulation study it was necessary to revise and extend the methods/software chosen. Moreover some smaller bug fixes were needed to make the software run in the first place. All web page reference mentioned were assessed and working on February 21, 2014.

[1] The Initiative for Integrative Psychiatric Research, http://ipsych.au.dk

2 Methods

2.1 G×E interaction analysis

With the intention to show the reasonableness of doing a more comprehensive study of a larger set of $G \times E$ methods we picked one version of the popular machine learning and data mining MDR method, model-based MDR (MB-MDR) (Calle et al., 2008), and one version of the machine learning method logic regression (Ruczinski, 2000; Kooperberg et al., 2001; Ruczinski et al., 2003), logic feature selection (logicFS) (Schwender et al., 2011). Furthermore we intend to compare these methods with a traditional two-step logistic regression with a step one consisting of choosing some relevant subset of the SNPs and a comprehensive search in this subset for interactions with the environmental variable in step two.

To check for main effects and two-way interactions in the simulated data we furthermore applied the *boosted one-step statistics* (BOSS) method by Voorman et al. (2012) and the *boolean operation-based screening and testing* (BOOST) method by Wan et al. (2010). All computations were carried out using the GenomeDK cluster[2].

Moreover, parameter estimates of the simulated models were investigated using logistic regression to check the performance of the simulation procedure. Calculations were carried out using R (R Core Team, 2013) if not mentioned otherwise.

2.1.1 Single step methods

BOSS has been implemented in the R package boss can be used for GWAS with repeated measures or related individuals, i.e. in situations with correlated errors. However we found that this is also an efficient method to search for G×E between each single SNP and a one environmental factor. This use of the software is not noted in Voorman et al. (2012) but there are option in boss to include an interacting variable.

2.1.2 Two-step methods

BOOST is a two-stage (screening and testing) that has been implemented as a commandline-based software[3] written in C and examines all pairwise SNP-SNP interactions very efficiently but only outputs (and test) results above a certain threshold. Along with the four degrees of freedom interaction results BOOST also outputs results from testing for association with single markers using two degrees of freedom, i.e. genotype-based test. The program output test statistics—not p-values.

Especially for models without main effects BOOST outperformed several other methods investigated in a review by Wang et al. (2011). In settings including main effects Wang et al. (2011) showed that the commandline-based C++ implemented *tree-based epistasis association mapping* (TEAM) by Zhang et al. (2010) was performing better but we have been unable to run TEAM[4] on the cluster.

[2] High-performance computing (HPC) environment established by the Genome Denmark project (http://genome.au.dk)

[3] http://bioinformatics.ust.hk/BOOST.html

[4] http://sourceforge.net/projects/epistasis/files/

For interactions of higher order than two we used BOSS and BOOST to choose an informed subset of the SNPs. We then carried out ”brute force” logistic regression analyses of either all interactions at the chosen level (e.g. all three-way interactions) or all interactions including a specific environmental exposure variable.

2.1.3 Model-Based MDR (MB-MDR)

The MDR approach was introduced by Ritchie et al. (2001) and have been applied with some success for detection of gene gene interaction. The method is nonparametric and do not make assumptions on the genetic penetrance model, i.e. it is model-free. MDR reduces the dimension of say a multilocus higher-order interaction by classifying each cell of the multi-dimensional space as either *high-risk* or *low-risk*, i.e. a one-dimensional high/low factor. For an overview of the procedures and a review of the classical version of MDR we refer to Motsinger et al. (2006). Many variants and improvements of this method have been developed (Moore et al., 2010; Pan et al., 2013).

We decided to use the MB-MDR proposed by Calle et al. (2008) which has been found to generally have higher power than MDR especially in situations with presence of genetic heterogeneity and phenocopies where MDR tends to have less success (Calle et al., 2008; Cattaert et al., 2011). The principal difference between MB-MDR and the classical MDR approach is that MB-MDR only merges genotype combinations that show significant evidence of high or low risk. The remainder, i.e. combinations with no evidence or insufficient sample size, are merged into a third category. The idea is to avoid noise from cells that are not important for the association effect. MB-MDR was first implemented and used for case-control studies, i.e. binary traits, but was later extended to quantitative traits (Mahachie John et al., 2011) and censored traits (Van Lishout et al., 2013).

The procedure of MB-MDR consist principally of three steps, see Figure 1 of (Cattaert et al., 2011). In step 1 all possible combinations of the k factors ($k = 1,2,\dots$) are tested for association with the trait. The choice of test depends on the trait type and may as such also be parametric or nonparametric. In step 2 the p-values for the test statistic calculated in step 1 are thresholded against some reference critical value p_c which per default is 0.1as recommended by Cattaert et al. (2011). Cells with $p < p_c$ are classified as high risk (H) or low risk (L) depending on the direction of the effect whereas cells with $p \geq p_c$ are classified as no risk evidence (O). For additional computational efficiency, cells with group size (e.g. cases+controls) less than a second threshold (default is 10) are also classified as O. In step 2 a second round of association tests is calculated for these HLO vectors and again the method allows for different testing strategies (Calle et al., 2008; Cattaert et al., 2011). In step 3 the significance of the test statistics from step 2 are determined with correction for multiple correlated tests using resampling-based step-down *maxT* adjusted p-values (Westfall et al., 1993).

For the calculations we used a C++ implementation[5] (Cattaert et al., 2011) which includes the efficient implementation of the multiple testing algorithm *MAXT* by Van Lishout et al. (2013). Also an even faster and so far unpublished algorithm (*speedMAXT*) is available and trades a slight higher false-positive rate for time (personal communication, François van Lishout, January 2014). The test statistic currently used by the programme is the maximum of the two tests H vs L/O and L vs H/O. The software includes the useful

[5] http://www.statgen.ulg.ac.be/software.html

opportunity to run parallel workflow both for the *MAXT* algorithm and to an even greater extent for the speedMAXT algorithm. The software handles interactions up to 3D, i.e. single markers, 2- or 3-way interactions, and can be used both for binary, continuous and time-to-event (censored) traits. Furthermore, it is possible to run for example all 3-way interactions between 1 (fixed) environmental factor and all possible SNP-SNP combinations which obviously is much faster than if all possible 3-way interactions had be search.

For the adjustment of main effects it is possible to choose co-dominant or additive (or none) coding of the genotypes. Mahachie John et al. (2012) recommends to always account for main effects of the SNPs under investigation for interaction and to do this as an integrated part of using MB-MDR as this adequately controls false positive findings. Furthermore they concluded that the co-dominant correction should be preferred as the additive coding may be insufficient and lead to overly optimistic results (c.f. Mahachie John et al., 2012).

2.1.4 Logic Feature Selection (logicFS)

Logic feature selection is a variant of the machine learning method called logic regression (Ruczinski, 2000) proposed by Schwender et al. (2008). It can be used to detect and quantify importance of genetic interactions in e.g. case-control studies by use a simulated annealing search algorithm within a regression framework including e.g. linear regression, logistic regression and Cox regression as possible responses. It is a prerequisite that the predictors are either binary (0/1, yes/no etc.) or can be formulated as a Boolean combination of binary variables. Thus continuous variables can only be used after categorisation though they may also enter as covariates in the regression. Under these considerations, environmental variables can be included and importance of $G \times E$ interactions may be investigated.

The method is implemented in the R Bioconductor[6] package logicFS and uses a so-called *bagging* approach (Breiman, 1996; Schwender et al., 2008) to stabilise the search for interactions. This is done by applying the simulated annealing algorithm multiple times to B bootstrap samples drawn by sampling with replacement from the original data. The original observations which are not in a given bootstrap sample are referred to as the out-of-bag (oob) observations. These are used to determine the importance of prime implicants (variables and interactions included in the proposed models) in terms of a *variable importance measure* (VIM), see Schwender et al. (2008) and the Appendix. During this process each *logic expression* (tree) from the resulting logic model is turned into a disjunctive normal form (DNF) which is an OR-combination of AND-combinations (the prime implicants).

2.2 Simulation of data

Case-control samples with individual-based genotypic data were generated using the Python[7] -based simulation environment simuPOP[8] (Peng et al., 2005; Peng et al., 2012). We used our own collection of scripts obtained by modification of a number of publicly

[6] http://www.bioconductor.org

[7] www.python.org

[8] http://simupop.sourceforge.net/

available simuPOP-based scripts. More details are given in the Appendix but in short, case-control samples are generated by use of the rejection sampling algorithm devised by Peng et al. (2010) and procedures from the Gene-Environment iNteraction Simulator (GENS) (Amato et al., 2010) and GENS version 2[9] (GENS2) (Pinelli et al., 2012) to control the penetrance while allowing for G×E interaction between two disease predisposing loci (DPLs) and one disease predisposing environmental variable (DPE).

Simulations are specified by the following parameters: expected disease prevalence in the population (m), the name (id) of DPLs (one or two), relative risk (RR) of the high risk homozygote compared with the low risk homozygote (expected risk ratio), a dominance parameter ($W \in [0,1]$), and parameters of the environmental variable plus the effect in terms of odds ratio (OR) of a one-unit increase in the environmental exposure for the (two-locus) genotype conferring the highest risk. The dominance parameter determines the relative risk of the heterozygote genotype as RR^W. Furthermore models for G×G and $G \times E$ needs to be specified.

3 Results

3.1 Generation of case-control samples

3.1.1 Choice of initial sample and generation of base sample

To ensure realistic correlation between the SNPs we used phased genomic data from the HapMap3 (International HapMap 3 Consortium, 2010) database[10] for the initial population consisting of the 993 unrelated subjects obtained by merging all 11 HapMap3 populations. Despite of this initial mix of populations, we treated this as a single population and migration was not considered. Peng et al. (2010) also used this strategy and showed that it works reasonably well. The reason for using the combined sample was to avoid bottleneck effects that may result from small founder population that are rapidly expanded. Moreover, even though the sudden population admixture caused long-range admixture LD this decayed enough over generations that long-range LD was not observed in the final simulated population (Peng et al., 2010). That LD is broken down relatively fast in expanding populations was also shown by Slatkin (1994).

Using a Wright-Fisher forward-time simulation with mutation and recombination, the initial population was then expanded linearly for 500 non-overlapping generations to a *base population* of 50,000 individuals. The effective population size is then approximately the harmonic mean of the census sizes in individual generations (Wright, 1938; Crow et al., 1970), and thus equals N_e=12,658. This appears to be a reasonably effective size for the present human population. The effective population size used during phasing of HapMap 3 release 2 for CEU was 11,418 according to the document *hapmap3_r2_phasing_summary.doc* available from the HapMap3 FTP site[10].

[9] http://sourceforge.net/projects/gensim/

[10] ftp:ftp.ncbi.nlm.nih.govhapmapphasing2009-02_phaseIIIHapMap3_r2

3.1.2 Choosing chromosomal regions and SNPs

For the large scale simulation study genome-wide strategies for G×E interaction analysis is the goal and for that purpose we would choose all common SNPs from the HapMap3 populations that are present on a specific commercial GWAS SNP chip. However, we start out by considering a smaller set of SNPs. Although the exact selection of loci do not matter for simulation purposes, we decided to mimic a strategy that might also be used to reduce the number of markers investigated for interactions in a real study. We simply selected the putative genetic linkage regions for schizophrenia in Caucasians based on a meta-analysis by Ng et al. (2009)[11]: 2q33.3–36.3 (206.3–228.3 Mb), 3p14.1–q13.32 (71.6–120.2 Mb), 5q31.3–35.1 (141.8–167.7 Mb), 6p21.31–12.1 (33.9–56.6 Mb), 8p22–12 (15.7–32.7 Mb), and 16p13.12–q12.2 (13.2–51.5 Mb). An updated list including more recent findings such as those from the Psychiatric Genomic Consortium (PGC) (Ripke et al., 2011; Ripke et al., 2013) could have been applied. From these six regions we chose SNPs that was present on the Illumina HumanHap550 chip and common to all 11 HapMap3 populations. We added a buffer zone of 10% of the region size to the ends of each region to allow an unrestricted choice of DPLs from the putative regions without risking edge effects. Alternatively, we could have excluded the outermost 10% or so of the region when choosing these variants. A total of 35680 of the 561,466 (547,458 autosomal) SNPs present on the HumanHap550 chip were positioned in these regions. Another restriction that should be noted is the assumption that the two DPLs are not in LD. To ensure this we simply chose DPLs from two different chromosomes.

At first we restricted to SNPs with MAF>0.05 in the base population leaving 33,053 SNPs for analyses (Foldager et al., 2013). Due to numerical problems in the optimiser that converts population features to model parameters (in the MLM) we subsequently included only SNPs with a genotype frequency of at least 0.05 for the minor allele homozygote. The problem was most likely caused by cells (combinations of SNPs and/or environmental factors) with low or zero counts: under random mating and thus Hardy-Weinberg proportions, a MAF above 0.05 only corresponds to a genotype frequency above 0.0025. This might be a valid approach for single-marker analyses but for G×G and G×E interaction analyses it becomes a numeric problem. After this more stringent inclusion criteria 19,097 SNPs remained in the regions: 2719 in chromosome (chr) 2, 4495 in chr 3, 3341 in chr 5, 2726 in chr 6, 2899 in chr 8, and 2917 in chr 16. Finally, to speed up calculations further, we decided to use only SNPs from the two chromosomal regions that included the DPLs.

3.1.3 Generation of case-control samples

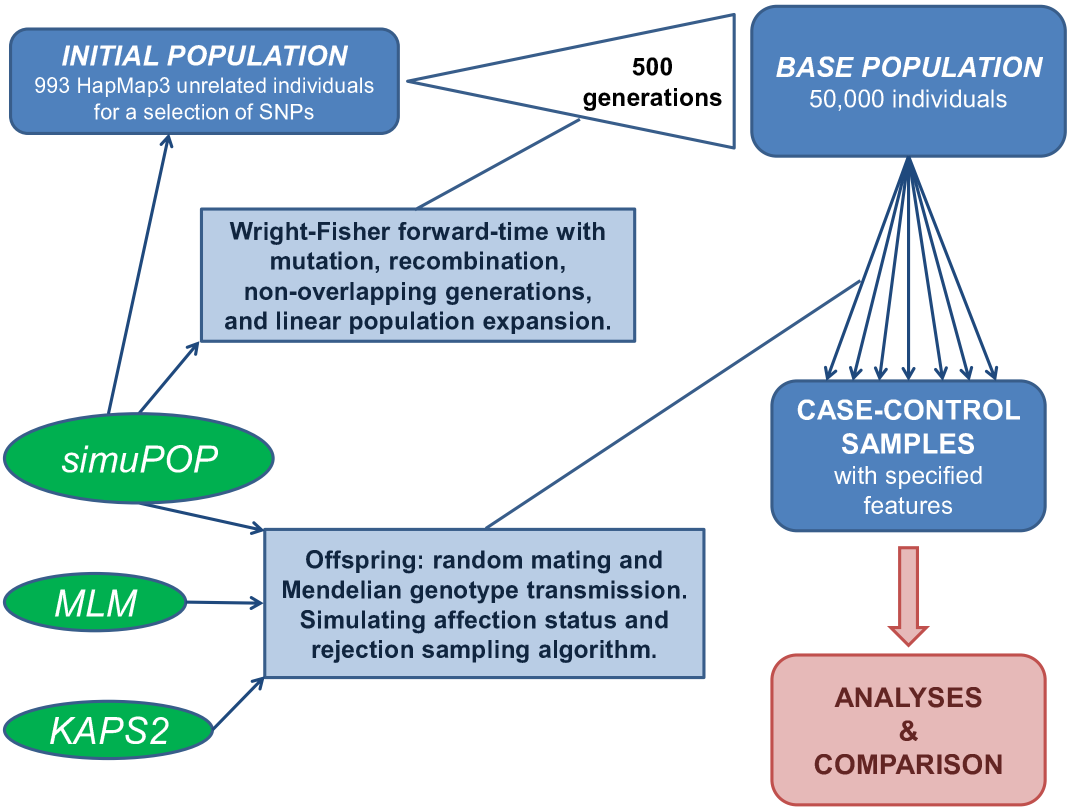
We decided to simulate from all sixteen possible combinations of two minor allele frequencies ($MAF \in \{0.3, 0.4\}$) and two risk ratios ($RR \in \{1, 1.2\}$) between high risk and low risk homozygote genotypes for each DPL, two odds ratios ($OR \in \{2, 5\}$) related to a one unit increase of the DPE, and two exposure proportions given by a probability parameter ($\pi \in \{0.25, 0.5\}$) of a binomial environmental distribution ($bi(1, \pi)$ distribution, i.e. a Bernoulli distribution). We currently only consider models with two DPLs and one DPE and we use the gene-environment interaction model referred to as *GEM* in Pinelli et al.

[11] http://www.szgene.org/linkage.asp

(2012) and no epistatic changes of the penetrances. No extra noise from non-predisposing environmental factors were included and we use the same parameters for both DPLs with $W = 1$ (i.e. a dominant genetic model). Furthermore, we use a fixed sample size of 10,000 with equally many affected and unaffected individuals and assumes the expected disease prevalence to be 1%. A flowchart of the simulation procedures is shown in Figure 1.

Figure 1

Flowchart for the G×E simulation study.


INITIAL POPULATION
993 HapMap3 unrelated individuals for a selection of SNPs
500 generations
BASE POPULATION
50,000 individuals
Wright-Fisher forward-time with mutation, recombination, non-overlapping generations, and linear population expansion.
simuPOP
MLM
KAPS2
Offspring: random mating and Mendelian genotype transmission. Simulating affection status and rejection sampling algorithm.
CASE-CONTROL SAMPLES
with specified features
ANALYSES & COMPARISON

3.2 Checking the simulated data

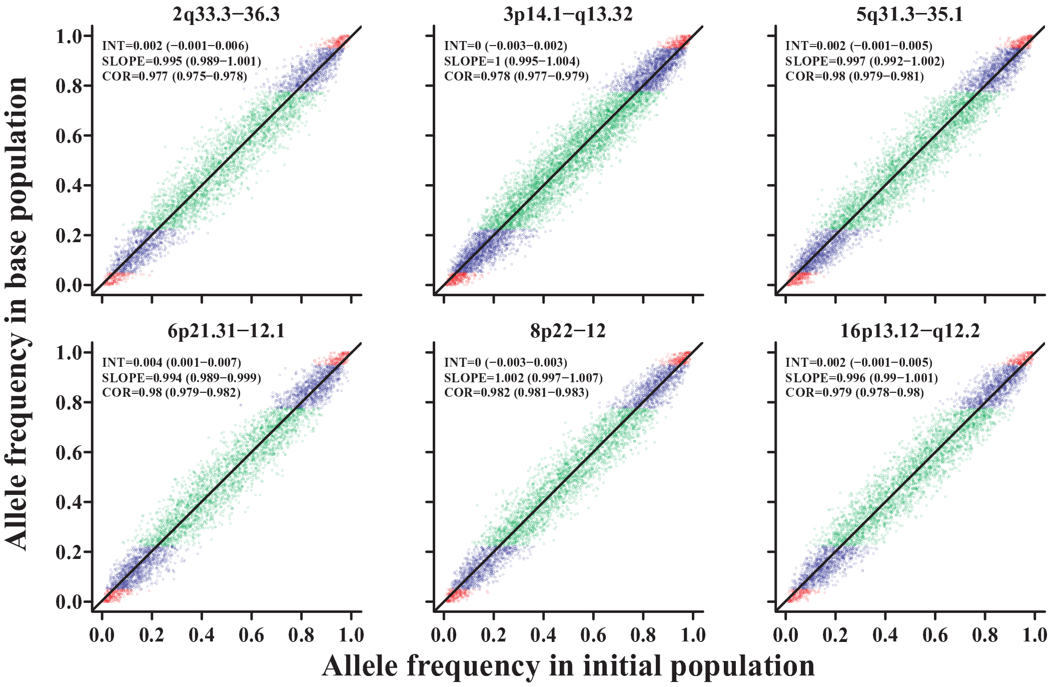
Table 1 shows the allele frequencies of the DPLs observed in the initial population, and the allele and genotype frequencies observed in the base population. The site frequency spectrum plots of allele frequencies in the base population versus the initial population for each of the 6 regions are shown in 2. The reduction given by excluding SNPs according to $P(m/m) > 0.05$ is indicated by green coloured points. The extra SNPs added by the use of the less stringent criterion MAF>0.05 are the blue points whereas red points are those exclude by both criteria. There is no indication of serious problems.

Table 1 DPL allele and genotype frequency

			Genotype frequency[a]		
ID (chr)	MAF_{init}	MAF_{base}	M/M	M/m	m/m
rs4257797 (5)	0.34	0.30	0.49	0.42	0.090
rs1781740 (6)	0.23	0.30	0.49	0.42	0.089
rs2941399 (6)	0.38	0.40	0.36	0.48	0.16
rs7000415 (8)	0.41	0.40	0.36	0.48	0.16

[a] M=major allele, m=minor allele

Figure 2



Site frequency spectrum. Green points are SNPs for which the frequency $P(m/m)$ of the minor homozygote is above 0.05, blue points are the extra points obtained using a limit of MAF>0.05, and red points are those excluded by both limits. That is, all SNPs is the union of green, blue and red points. The slope line is the least squares fit using all points.

3.2.1 BOSS and BOOST results

Summary statistics of p-values obtained using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios are presented here for: 1) single-marker $\chi^2(2)$ genotype-based tests from BOOST (Figure 3); 2) single-marker additive tests adjusted for DPE main effect (Wald tests) from BOSS (Figure 4); 3) two-way interaction tests adjusted for SNP and DPE main effects (Wald tests) from BOSS (Figure 5). Furthermore results from two-way SNP-SNP interaction tests adjusted for main effects ($\chi^2(4)$) from BOOST are summarized (Figure 6). The curves shown in the figures are smoothing splines[12] and all values are plotted on a minus-log-base-10 scale against chromosomal position (base pairs).

Figure 3 indicate that the simulated samples adhere to the simulated models with respect to genotypic main effects: RR=1.0 corresponding to no effect and RR=1.2 corresponding to a small main effect (same for both DPLs). There are no noticeable effects of varying the other parameters (MAF, OR and prevalence π of environmental exposure). It is noticeable that none of the minima are above the Bonferroni adjusted threshold in the scenarios where RR=1.0 whereas this is the case for a small set of markers when RR=1.2. Interestingly those above the threshold are not just markers in close proximity to the DPLs. Probably this is due to longer ranging LD. Note also, that the main effects (RR=1.2) would remain undetected in many of the samples when using the Bonferroni threshold.

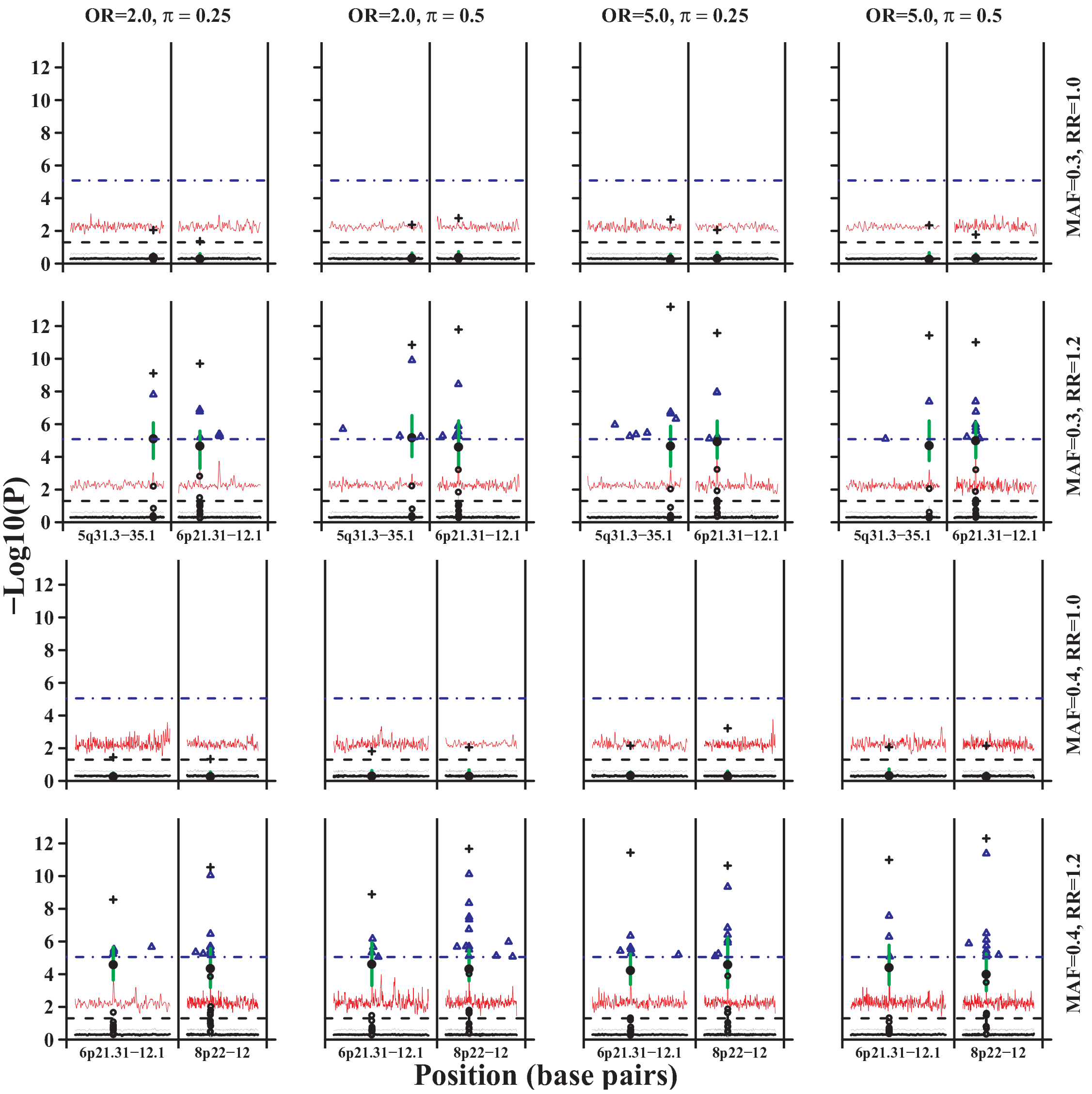
Figure 4 show that the adjusting for the environmental diminishes p-values of the genetic main effects. In accordance with the simulated models this reduction is larger for larger environmental effect, i.e. more pronounced when the samples were sampled with OR=5.0 than with OR=2.0. The other conclusions from BOOST single-marker tests (figure 3) remains.

Figure 5 show summary statistics from BOSS G×E tests (two-way interaction between SNPs and DPE) adjusted for main effects of SNP and DPE. Here we note that the two-way interactions between DPLs and DPE are highly significant when the main effect of DPE is smaller (OR=2.0) and less prevalent ($\pi = 0.25$). Both increased prevalence of the environmental factor (DPE) and increase of its disease predisposing effect (OR=5.0) diminishes the significance of the interaction term. The LD effects mentioned for single-marker BOOST results are still visible. We chose to use the same range of the y-axis as we used in figure 3 and figure 4 though this means that some of the points for the DPLs is outside the range. The two points missing for the scenario with MAF=0.3, RR=1.2, OR=2.0 and $\pi = 0.25$ are P=2.8e-18 and P=9.0e-18. The one point missing for the corresponding scenario with MAF=0.4 is P=8.6e-18.

Figure 6 shows bar plots summarising BOOST G×G $\chi^2(4)$ genotype-based tests (two-way interaction between SNPs). Only test statistics >30 (P<4.9e-6) were used and the tests are adjusted for main effects of the interacting SNPs. The bars are the number of samples (out of 100) where the SNP was present in at least one SNP-SNP interaction with a test statistic above the threshold of 30. In agreement with the models simulated (no epistasis), no systematic patterns are apparent and the DPLs are not more often part of G×G interactions than the other SNPs.

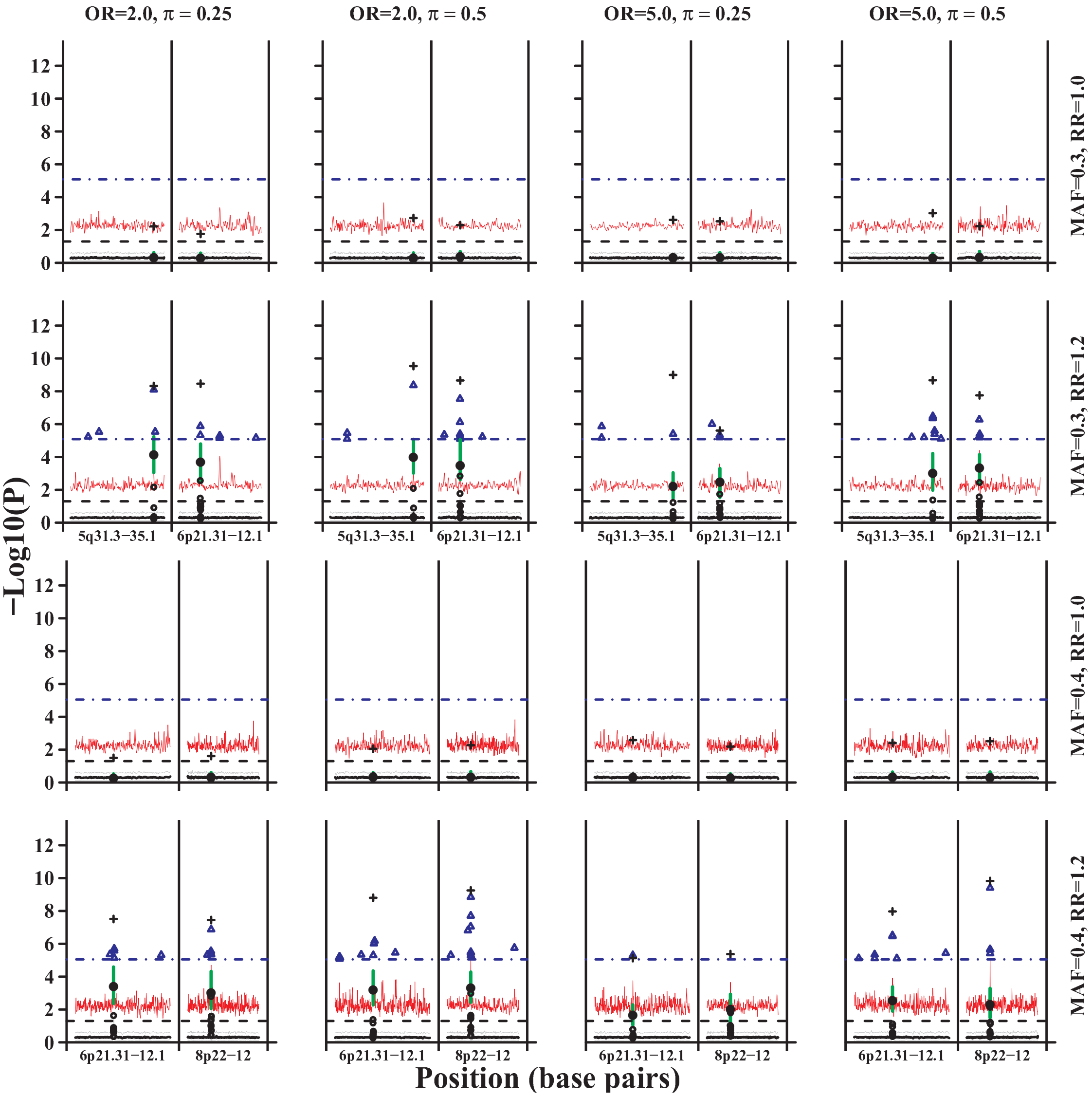
[12] R function smooth.spline with default options except for the curves showing minimum of p-values where all knots were used.

Figure 3



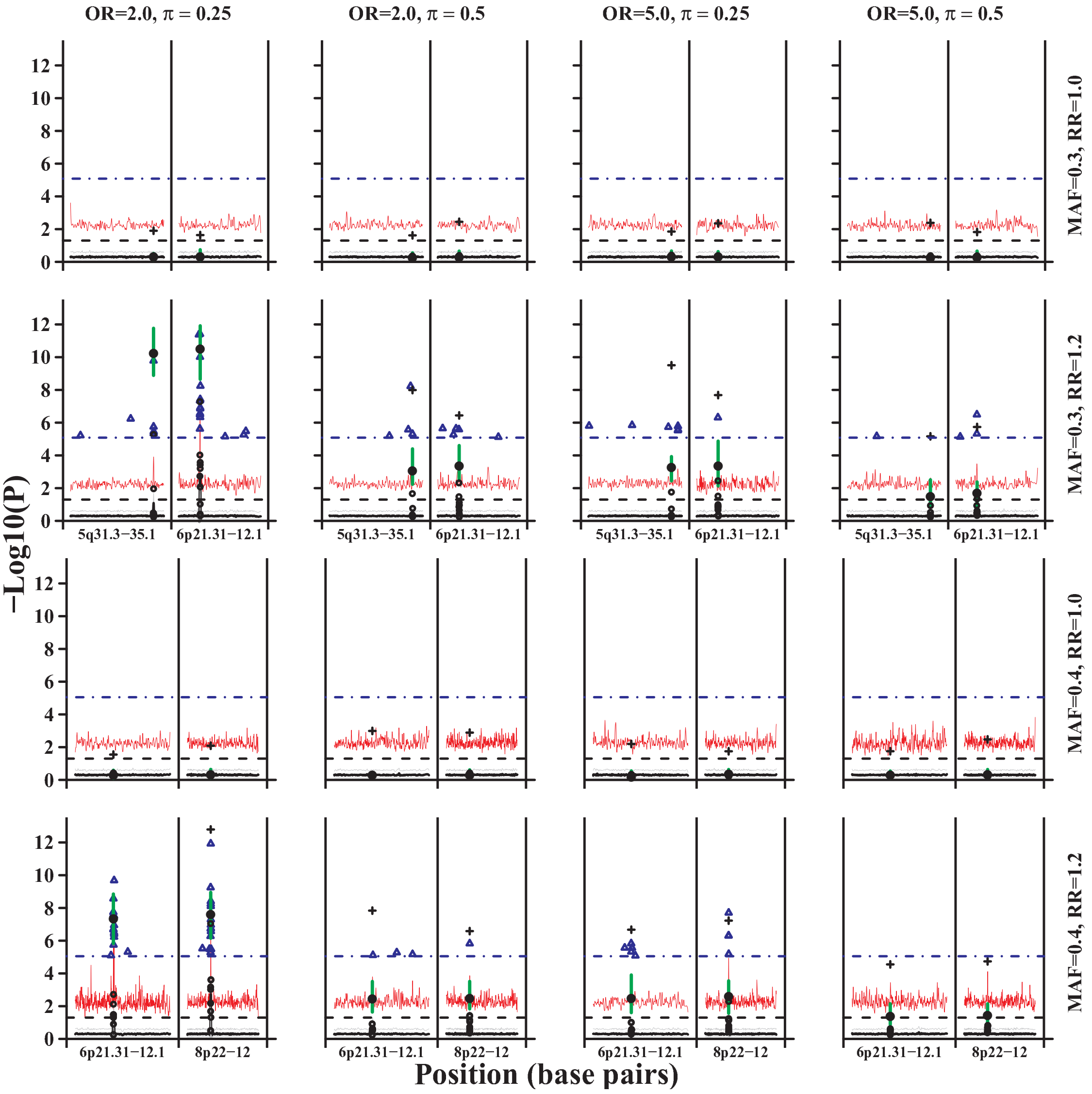
BOOST: Single-marker genotype-based test. The figure shows summary statistics of p-values from single-marker $\chi^2(2)$ genotype-based tests from BOOST using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios (parameters are indicated). Curves are smoothing splines[12] and all values are plotted on a minus-log-base-10 scale against chromosomal position (base pairs). The grey curves indicate upper and lower quartiles, the black curves are medians, and the red curves are minimum of p-values (maximum of $-\log_{10}(P)$) from the 100 samples at each marker. The green bars show the interquartile range for the DPLs. The black points are medians of p-values for DPLs and their each of 5 neighbouring markers at each side (those for DPLs are filled bullets and a bit larger). The '+' points are minimum of p-values for the DPLs. The horizontal broken lines indicate the nominal significance level (0.05) and the Bonferroni threshold (dash-dot blue line) after adjustment for all SNPs in the two regions, and the blue triangles show non-DPL minimum of p-values above this threshold.

Figure 4



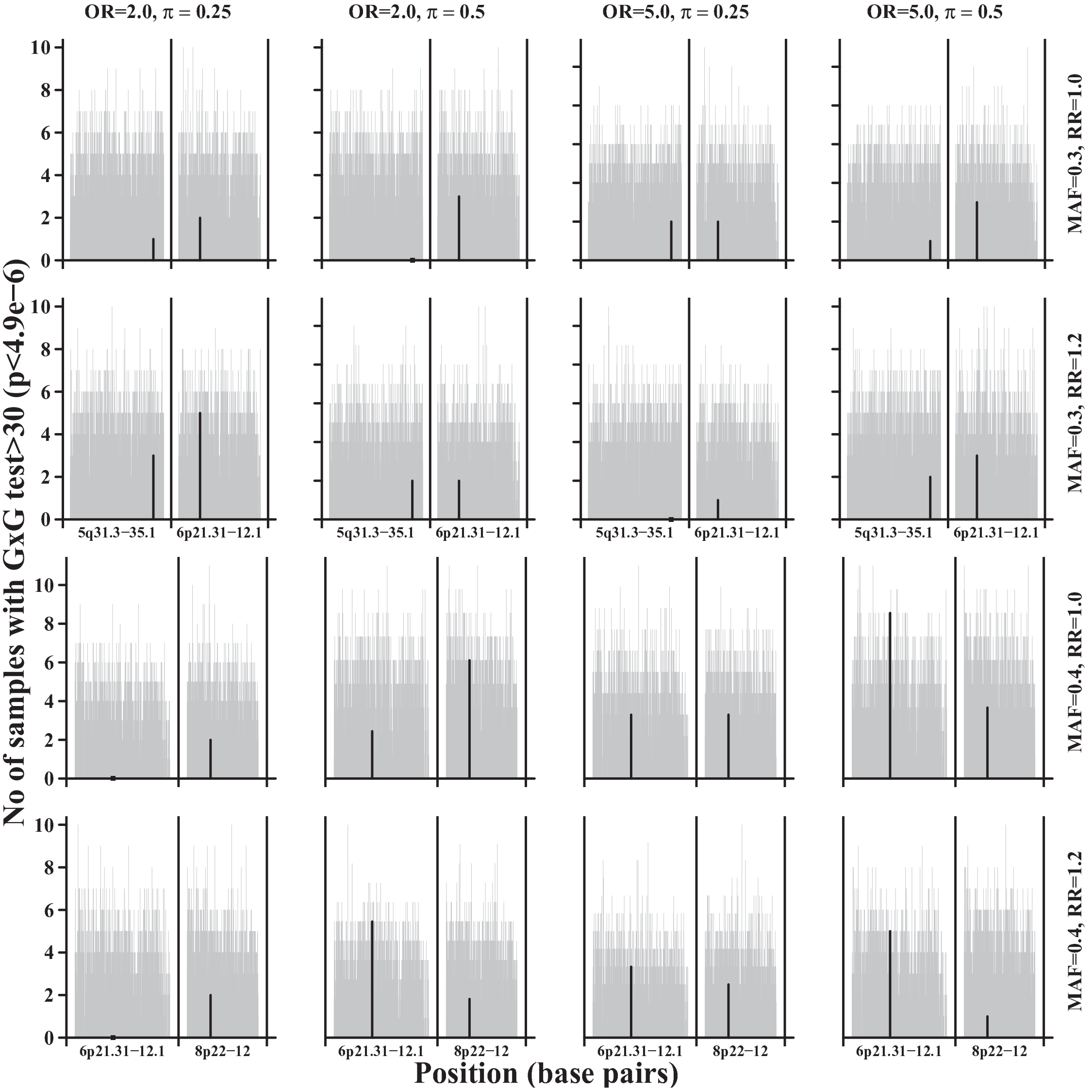
BOSS: Single-marker additive DPE adjusted test. The figure shows summary statistics of p-values from single-marker additive tests adjusted for DPE main effect (Wald tests) from BOSS using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios (parameters are indicated). Curves are smoothing splines[12] and all values are plotted on a minus-log-base-10 scale against chromosomal position (base pairs). The grey curves indicate upper and lower quartiles, the black curves are medians, and the red curves are minimum of p-values (maximum of $-\log_{10}(P)$) from the 100 samples at each marker. The green bars show the interquartile range for the DPLs. The black points are medians of p-values for DPLs and their each of 5 neighbouring markers at each side (those for DPLs are filled bullets and a bit larger). The '+' points are minimum of p-values for the DPLs. The horizontal broken lines indicate the nominal significance level (0.05) and the Bonferroni threshold (dash-dot blue line) after adjustment for all SNPs in the two regions, and the blue triangles show non-DPL minimum of p-values above this threshold.

Figure 5



BOSS: Two-way G×E interaction test. The figure shows summary statistics of p-values from two-way interaction tests adjusted for SNP and DPE main effects (Wald tests) from BOSS using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios (parameters are indicated). Curves are smoothing splines[12] and all values are plotted on a minus-log-base-10 scale against chromosomal position (base pairs). The grey curves indicate upper and lower quartiles, the black curves are medians, and the red curves are minimum of p-values (maximum of $-\log_{10}(P)$) from the 100 samples at each marker. The green bars show the interquartile range for the DPLs. The black points are medians of p-values for DPLs and their each of 5 neighbouring markers at each side (those for DPLs are filled bullets and a bit larger). The '+' points are minimum of p-values for the DPLs. The horizontal broken lines indicate the nominal significance level (0.05) and the Bonferroni threshold (dash-dot blue line) after adjustment for all SNPs in the two regions, and the blue triangles show non-DPL minimum of p-values above this threshold. A few values are smaller than 1e-13 and thus outside the range of the y-axis.

Figure 6



BOOST: Two-way G×G interaction test. Bar plots summarising results from two-way SNP-SNP interaction tests adjusted for main effects ($\chi^2(4)$) from BOOST using 100 simulated samples of 5,000 cases and 5,000 controls for each of the 16 scenarios (parameters are indicated). The bars show the number of samples each SNP appear in at least one G×G two-way interaction with the $\chi^2(4)$ statistic above a threshold of 30 (P<4.9e-6). The results for DPLs are shown with black bars and by a square point on the x-axis in cases where this count is zero.

4 Discussion

Finding and choosing the most appropriate method and software to simulated genetic data can be difficult but a web site was recently established to accommodate this process (Peng et al., 2013). Moreover, a thorough review of state of the art software for computer simulations of population and evolution genetics can be found in Hoban et al. (2012)—though *state of the art* changes quite rapidly in this field. We used the very general Python-based forward-time simulator simuPOP (Peng et al., 2005; Peng et al., 2012) which is able to simulate individuals with genotypes under many evolutionary scenarios.

In the present paper we simulated from sixteen scenarios but further scenarios will be considered for the full scale study by varying the sample size, varying the disease prevalence, varying the co-dominance parameter W (e.g. W=0.5 corresponding to additive effects and W=0 corresponding to recessive effects), combinations of MAF (i.e. not same MAF for both DPLs), inclusion of epistasis, inclusion of environmental noise, other values of RR, OR and π, other models and other restrictions on the SNPs included. It would be desirable to be able to also model over-dominance of the heterozygote genotype and to have a more precise control on the epistatic parameters. We might also consider simulating with uneven proportions of cases and controls. It is possible to also use non-neutral selection, migration, non-overlapping generations and non-linear expansion for the evolution of the initial population, thus providing more realistic scenarios, but making conduction and interpretation more complicated.

Pinelli et al. (2012) noted that there is a balance between restrictions and the demands for the user of the software to choose parameter values. To make their system more user friendly and available maybe to a broader audience constraints were therefore imposed to reduce the complexity of the systems in GENS2. This should also be a way to ensure that the user knows exactly what was simulated and thus what the methods tested should be able to find. However, we find that user have very little handle on the changes imposed by the epistasis option as most parts of that matrix is determined by an optimisation algorithm. Furthermore, in our opinion, some of the restrictions are not too obviously meaningful. Especially the omission of main effects in the interaction model and restriction away from over-dominance of the heterozygote genotype may not always be natural.

The use of an greatly admixed initial sample induce long-range admixture LD (Smith et al., 2005) but due to the many subsequent generations in an expanding population and thereby recombinations, this LD have been reduced (Slatkin, 1994) enough that Peng et al. (2010) found no sign of elevated long-range LD in the simulated populations. The reason for using the combined samples were to reduce the problem of bottleneck effects (and associated genetic drift) that may result from small founder population being rapidly expanded. Peng et al. (2010) expect the availability of large sequenced samples of high coverage will enable the possibility to obtain population-specific samples, i.e. to simulate from population-specific initial populations rather than from a mixture of populations. This is, however, unlikely to affect the results of the current simulation study.

Half of the scenarios have been run using the *speedMAXT* parallel workflow of MB-MDR to search for three-way interactions consisting of the DPE and all combinations of SNP pairs. The options were set such that only the 1000 combinations with the largest test statistic were stored and evaluated. However, from a first quick sneak peek it seems that these 1000 test statistics are almost of the same size and thus apparently not revealing the simulated interaction. This is something we have to look closer into and thus, at this time, no MB-MDR results will be presented. No analyses using logicFS have yet been completed.

4.1 Comparisons of $G \times E$ interaction methods

Complicated problems tend to yield complicated answers and it is not straightforward to compare complex approaches like machine learning methods where the performance not only depend on the often complicated problems they are applied to, but also to some extent depends on the user's ability to tune parameters of the algorithms. Therefore, when applying machine learning methods in practice several different methods, algorithms and/or sets of parameters are often used. So far we have just used default settings if possible or reasonable but a more thorough investigation should be made for choosing algorithmic parameters in such a way that the comparisons are reasonably fair. Some preliminary considerations of how to compare and measure the performance and the G×E methods are given here.

Computational speed and scalability is one issue that may be important to consider. To be realistically useful for genome-wide interaction studies the methods and software should probably be parallelisable to as large an extent as possible and this may well be a major factor to consider when choosing methods. Measuring the computing time may in itself be problematic and physical limitations of the machine/system may also affect methods differently. Limits in memory may slow some methods but not others, harddisk speed may affect some methods more than others, processor speed likewise. Also, there may well be a trade-off also between speed and precision. An example of an examination of speed versus sensitivity can be found in Brinza et al. (2010).

Many methods are only implemented to detect two-way interactions whereas others search for higher order interactions too. If the simulated (known) truth only implies a two-way interaction should we then prevent the algorithms from searching for interactions of higher order? And how about methods that allow for multiple interacting factors (even if restricted to two-ways) should these then be confined to one interaction term, e.g. a maximum of one tree? Furthermore, some methods only allows for a somewhat limited number of factors (e.g. SNPs) to analyzed whereas others searches through all possible SNP-SNP interactions in a GWAS and others again uses a multi-step approach to shrink the search space. That is, how do we handle differences in complexities between methods when we compare them? Along these lines, some methods may include more terms than needed to capture the true association. As an example, if we set the limit of the model size higher than needed then many machine learning methods are likely to be overfitting but might still also find the true signal.

Concerning the ability to predict the outcome we may consider using the misclassification rate (MCR), the (predictive) accuracy (AC), the balanced accuracy (BA) and the kind. In Table 2, the so-called confusion matrix, we have shown some of the terms often used in this respect. Sometimes cross validation is used to assess and compare the prediction ability of different methods but we may just as well utilise that we have simulated multiple datasets from the same population. If we have fitted a model using one of the 100 datasets (training data) then we can access the fitted models ability to predict affection status in the other 99 datasets (test data).

Table 2 The confusion matrix

		Observed		
		Affected	Unaffected	Total
Predicted	Affected	a (TP)	c (FP)	a+c
	Unaffected	b (FN)	d (TN)	b+d
	Total	a+b	c+d	N

TP: true positive; FP: false positive; FN: false negative; TN: true negative; N=a+b+c+d=TP+FN+FP+TN; sensitivity=a(a+b); precision=a(a+c); specificity $= d/(c+d)$; negative predictive value (NPV)=d(b+d)

Let $\mathbf{y}^* = (y_1^*, \ldots, y_N^*)$ denote the observed binary response for a test dataset and let $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$ be that predicted by a model fitted on the training data but evaluated on the observed p predictors X ($N \times p$ matrix) from this test dataset. Then the squared difference $(y_i^* - \hat{y}_i)^2$ will be 1 if the i'th individual is misclassified by the fitted model and the misclassification rate may therefore be calculated as

$$MCR = \frac{1}{m}\sum_{i=1}^{m}(y_i^* - \hat{y}_i)^2 \tag{1}$$

This is also referred to as the mean model error rate in Wolf et al. (2010) and in the case of a continuous response it is the mean squared error (MSE). Comparing with the entrances of the confusion matrix (Table 2) we see that the MCR = (FP+FN)/N. The (predictive) accuracy is AC=(TP+TN)/N and since N=(TP+FP+TN+FN) we see immediately that AC=1-MCR. Thus it is a matter of taste if AC or MCR is used. In case of imbalanced datasets (unequal number of cases and controls) some advocate using the balanced accuracy (BA) defined as the average of the sensitivity and specificity, i.e. BA=(sensitivity+specificity)/2.

Finally, it is standard to compare methods by their power (sensitivity) and ability to maintain the level of significance (type I error rate = rate of false positive = 1-specificity). The power can be determined as the proportion of simulated test datasets in which the disease associated effects are detected by the method. But what if some of the effects are detected? And what if main effects are detected but not the interactions? One also need to consider how indirect association count, i.e. association with markers that are in LD with the causal variant. Measures of importance like the VIM in logicFS may be a way to handle this and it should be investigated if similar measures exists or can be developed for the other methods to be compared. To determine maintenance of specificity, simulations without disease association (null models) are usually made and the proportion runs with erroneously detection of an effect then determines the type I error rate. But if interactions are the point of interest what is then the correct null model? Should it be one with or without main effects? This is the same problem that makes permutation-based testing difficult to implement for methods searching for interactions.

4.2 Concluding remarks

We initiated analyses using tradition two-step logistic regression and the two machine learning/data mining methods *logicFS* and *MB-MDR*. A practical issue in connection with machine learning methods is the need of complete data (no missing genotypes or other measures) which is not needed when using logistic regression or other generalised linear models. The MB-MDR and logicFS softwares are not exceptions from that rule. Another problem is scalability of the methods with limits on the the number of markers and/or other factors/covariates that can be included. These limits may be software specific in terms of restrictions defined in the programs or hardware induced by memory limits or processor capacities. On the other hand, one of the advantages of using machine learning methods is the possibility to search for higher order interactions without being compromised by the need to adjust for multiple testing adjustment to a degree that the effect sizes or sample size have to be unrealistically large.

Finally, optimal methods should take genotype probabilities and thereby allow for imprecision (or variation) of genotyping (and/or imputation) as well as avoiding the need for complete data. This may be worth having in mind when choosing further G×E methods for comparison.

Appendix

simuGEMS

Applying the simuPOP simulation environment (Peng et al., 2005; Peng et al., 2012), we generated case-controls samples with individual-based genotypic data. We constructed our own set of Python scripts, borrowing massively from the available scripts[8] described in (Peng et al., 2010; Peng et al., 2012) and from GENS2[9] (Pinelli et al., 2012). We call this collection of scripts the simuPOP-based Gene-Environment Model Simulator (simuGEMS) and it is available from the corresponding author on request.

We simulated without selection (e.i. only neutral processes) on all SNPs and instead of using trajectory sampling to ensure specific MAF of predefined DPLs we picked at random among SNPs having a MAF of a certain size in what we refer to as the *base population* (see Results). Mutation rate and recombination intensity were both set at $1e-8$ and the recombination rate is then the intensity multiplied by the physical distance in basepairs (bp) between adjacent loci.

To generate case-control samples we implemented the rejection sampling method described in Peng et al. (2010). This rejection sampling algorithm is needed when considering diseases of low prevalence as the proportion of affected will usually be too low for random sampling from the base population to be feasible.

Affection status was generated by modified scripts from GENS2 to control the penetrance while allowing for gene-environment (G×E) interaction between up to two disease predisposing loci DPLs and one DPE, with the possibility to also introduce epistasis (G×G interaction) as a source of complexity. In addition to some bug fixes we extended with binomial (including binary) and multinomial DPE distributions.

It is as such no problem to simulate log-linear G×E models directly using simuPOP, see e.g. example 2 in Peng et al. (2010). But the set of parameters specifying the model have to be chosen in a way that the expected disease prevalence (found by integration over the sample space of the covariates) equals the assumed proportion of the population (e.g. 1%) given the generated allele or genotype distribution of the DPLs and under the selected distribution of the DPE. This is not trivial but it is what the GENS framework was designed to solve.

The calculation of penetrances in the interaction models is carried out by use of the mathematical approach Multi-Logistic Model (MLM) suggested by Amato et al. (2010). Using the Knowledge Aided Parameterization System (KAPS) (Amato et al., 2010) for one DPL one DPE or KAPS version 2 (KAPS2) (Pinelli et al., 2012) for two DPLs one DPE, values of biological and epidemiological parameters are translated to the coefficients of the MLM which corresponds in essence to penetrances of the various combinations of DPL (multilocus) genotypes as a function of the DPE.

The parameters consists of the expected disease prevalence of the sample (m), the name (id) of DPLs (one or two), relative risk (RR) of the high risk homozygote compared with the low risk homozygote (expected risk ratio), a dominance parameter ($W \in [0,1]$) , and parameters of the environmental variable plus the effect in terms of odds ratio (OR) of a one-unit increase in the environmental exposure for the (two-locus) genotype conferring the highest risk. The dominance parameter determines the relative risk of the heterozygote genotype as RR^W so that $W = 0$ corresponds to a dominant model, $W = 1$ is a recessive model, and otherwise a co-dominant model is obtained. Over-dominance ($W > 1$) cannot be modeled at present. The genotype frequencies of each DPL is calculated from allele frequencies under the assumption of random mating in non-overlapping generations, i.e. Hardy-Weinberg proportions. Furthermore models for G×G and $G \times E$ needs to be specified and chosen, see Pinelli et al. (2012).

VIM measures in logicFS

Two VIM measures are available in logicFS—one for single-tree models and one for the multiple-tree case. The latter may also be used for a single-tree model. In both cases a large positive value corresponds to a high importance of the prime implicant, a value close to zero indicates no importance, and a negative value shows that the prime implicant is obstructive for the classification (c.f. Schwender et al., 2008). Let $\mathscr{H}_b$ denote the set of prime implicants identified in the model fitted using the b'th bootstrap sample and let N_b denote the count of oob observations correctly classified by this model, $b = 1,\dots,B$. Let $\mathscr{H} = \{\mathscr{H}_b\}$ denote the set of all prime implicants observed in the B DNFs.

Single-tree VIM measure For the single-tree situation a VIM can be calculated for each element of $h \in \mathscr{H}_b$ in the following way: for $b \in 1,\dots,B : h \in \mathscr{H}_b$ remove this prime implicant from the DNF and re-count the number $N_b^{(-h)}$ of correctly classified oob observations using this reduced model. Correspondingly, $b \in 1,\dots,B : h \notin \mathscr{H}_b$ add this prime implicant to the DNF and re-count the number $N_b^{(+h)}$ of correctly classified oob observations using this extended model. Now the importance of h can be measured by (Schwender et al., 2008, equation 4.1):

$$VIM_S(h) = \frac{1}{B}\left(\begin{array}{l} \sum_{b:h\in\mathscr{H}_b}(N_b - N_b^{(-h)}) \\ +\sum_{b:h\notin\mathscr{H}_b}(N_b^{(+h)} - N_b) \end{array}\right) \tag{2}$$

Multiple-tree VIM measure In the case with more than one tree it is only possible to consider the part where prime implicants are removed. The prime implicant will be removed at once from all trees (models). We will stick to the same notation as for single-trees and let N_b^{-h} denote the number of correctly classified oob observations in the reduced model. The importance is then measured by (Schwender et al., 2008, equation 4.2):

$$\begin{aligned} VIM_M(h) &= \tfrac{1}{B}\textstyle\sum_{b=1}^{B}(N_b - N_b^{(-h)}) \\ &= \textstyle\sum_{b:h\in\mathscr{H}_b}(N_b - N_b^{(-h)}) \end{aligned} \tag{3}$$

References

Amato, R., Pinelli, M., D'Andrea, D. et al. (2010). 'A novel approach to simulate gene-environment interactions in complex diseases'. *BMC Bioinformatics* **11**: 8.

Borglum, A.D., Demontis, D., Grove, J. et al. (2013). 'Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci'. *Mol. Psychiatry*, Epub ahead of print.

Breiman, L. (1996). 'Bagging predictors'. *Machine Learning* **24**: 123–140.

Breiman, L. (2001). 'Random Forests'. *Machine Learning* **45**: 5–32.

Brinza, D., Schultz, M., Tesler, G. and Bafna, V. (2010). 'RAPID detection of gene-gene interactions in genome-wide association studies'. *Bioinformatics* **26**: 2856–2862.

Calle, M.L., Urrea, V., Malats, N. and Van Steen, K. (2008). *MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data*. Tech. rep. Universitat de Vic, 1–14.

Cattaert, T., Calle, M.L., Dudek, S.M. et al. (2011). 'Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise'. *Ann. Hum. Genet.* **75**: 78–89.

Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row,

Foldager, L., Als, T.D. and Grove, J. (2013). 'Comparison of methods for genome-wide gene-environment interaction analysis'. In: *Abstract book for XXI World Congress of Psychiatric Genetics (WCPG)*. International Society of Psychiatric Genetics (ISPG). Boston, MA, USA, 279–280.

Hoban, S., Bertorelle, G. and Gaggiotti, O.E. (2012). 'Computer simulations: tools for population and evolutionary genetics'. *Nat. Rev. Genet.* **13**: 110–122.

International HapMap 3 Consortium (2010). 'Integrating common and rare genetic variation in diverse human populations'. *Nature* **467**: 52–58.

Kooperberg, C., Ruczinski, I., LeBlanc, M.L. and Hsu, L. (2001). 'Sequence analysis using logic regression'. *Genet. Epidemiol.* **21 Suppl 1**: S626–S631.

Mahachie John, J.M., Cattaert, T., Van Lishout, F. et al. (2012). 'Lower-order effects adjustment in quantitative traits model-based multifactor dimensionality reduction'. *PLoS One* **7**: e29594.

Mahachie John, J.M., Van Lishout, F. and Van Steen, K. (2011). 'Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data'. *Eur. J. Hum. Genet.* **19**: 696–703.

Moore, J.H., Asselbergs, F.W. and Williams, S.M. (2010). 'Bioinformatics challenges for genome-wide association studies'. *Bioinformatics* **26**: 445–455.

Motsinger, A.A. and Ritchie, M.D. (2006). 'Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies'. *Hum. Genomics* **2**: 318–328.

Motsinger-Reif, A.A. and Ritchie, M.D. (2008). 'Neural networks for Genetic Epidemiology: past, present, and future'. *BioData Min.* **1**: 3–.

Murcray, C.E., Lewinger, J.P. and Gauderman, W.J. (2009). 'Gene-environment interaction in genome-wide association studies'. *Am. J. Epidemiol.* **169**: 219–226.

Ng, M.Y., Levinson, D.F., Faraone, S.V. et al. (2009). 'Meta-analysis of 32 genome-wide linkage studies of schizophrenia'. *Mol. Psychiatry* **14**: 774–785.

Pan, Q., Hu, T. and Moore, J.H. (2013). 'Epistasis, complexity, and multifactor dimensionality reduction'. *Methods Mol. Biol.* **1019**: 465–477.

Peng, B. and Amos, C.I. (2010). 'Forward-time simulation of realistic samples for genome-wide association studies'. *BMC Bioinformatics* **11**: 442.

Peng, B., Chen, H.S., Mechanic, L.E. et al. (2013). 'Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators'. *Bioinformatics* **29**: 1101–1102.

Peng, B. and Kimmel, M. (2005). 'simuPOP: a forward-time population genetics simulation environment'. *Bioinformatics* **21**: 3686–3687.

Peng, B., Kimmel, M. and Amos, C.I. (2012). *Forward-Time Population Genetics Simulations: Methods, Implementation, and Applications*. New Jersey, USA: Wiley-Blackwell,

Pinelli, M., Scala, G., Amato, R. et al. (2012). 'Simulating gene-gene and gene-environment interactions in complex diseases: Gene-Environment iNteraction Simulator 2'. *BMC Bioinformatics* **13**: 132.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Ripke, S., O'Dushlaine, C., Chambert, K. et al. (2013). 'Genome-wide association analysis identifies 13 new risk loci for schizophrenia'. *Nat. Genet.* **45**: 1150–1159.

Ripke, S., Sanders, A.R., Kendler, K.S. et al. (2011). 'Genome-wide association study identifies five new schizophrenia loci'. *Nat. Genet.* **43**: 969–976.

Ritchie, M.D., Hahn, L.W., Roodi, N. et al. (2001). 'Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer'. *Am. J. Hum. Genet.* **69**: 138–147.

Ruczinski, I. (2000). 'Logic regression and statistical issues related to the protein folding problem'. PhD thesis. Seattle: University of Washington, Dept. of Statistics.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003). 'Logic regression'. *J. Comput. Graph. Stat.* **12**: 475–511.

Schwender, H. and Ickstadt, K. (2008). 'Identification of SNP interactions using logic regression'. *Biostatistics* **9**: 187–198.

Schwender, H., Ruczinski, I. and Ickstadt, K. (2011). 'Testing SNPs and sets of SNPs for importance in association studies'. *Biostatistics* **12**: 18–32.

Slatkin, M. (1994). 'Linkage disequilibrium in growing and stable populations'. *Genetics* **137**: 331–336.

Smith, M.W. and O'Brien, S.J. (2005). 'Mapping by admixture linkage disequilibrium: advances, limitations and guidelines'. *Nat. Rev. Genet.* **6**: 623–632.

Van Lishout, F., Mahachie John, J.M., Gusareva, E.S. et al. (2013). 'An efficient algorithm to perform multiple testing in epistasis screening'. *BMC Bioinformatics* **14**: 138.

Voorman, A., Rice, K. and Lumley, T. (2012). 'Fast computation for genome-wide association studies using boosted one-step statistics'. *Bioinformatics* **28**: 1818–1822.

Wan, X., Yang, C., Yang, Q. et al. (2010). 'BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies'. *Am. J. Hum. Genet.* **87**: 325–340.

Wang, Y., Liu, G., Feng, M. and Wong, L. (2011). 'An empirical comparison of several recent epistatic interaction detection methods'. *Bioinformatics* **27**: 2936–2943.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons,

Wolf, B.J., Hill, E.G. and Slate, E.H. (2010). 'Logic Forest: an ensemble classifier for discovering logical combinations of binary markers'. *Bioinformatics* **26**: 2183–2189.

Wright, S. (1938). 'Size of population and breeding structure in relation to evolution'. *Science* **87**: 430–431.

Zhang, X., Huang, S., Zou, F. and Wang, W. (2010). 'TEAM: efficient two-locus epistasis tests in human genome-wide association study'. *Bioinformatics* **26**: i217–i227.

6.4 Paper 4[33]

Journal of Affective Disorders 148 (2013) 291–298

Contents lists available at SciVerse ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research report

An association study of suicide and candidate genes in the serotonergic system

Henriette N. Buttenschøn[a,b,*,1], Tracey J. Flint[a,1], Leslie Foldager[a,b,c], Ping Qin[d], Søren Christoffersen[e], Nikolaj F. Hansen[f], Ingrid B. Kristensen[g], Preben B. Mortensen[b,d], Anders D. Børglum[a,b,h], Ole Mors[a,b]

[a] Centre for Psychiatric Research, Aarhus University Hospital, Risskov, Denmark
[b] The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark
[c] Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[d] The National Centre for Register-based Research, Aarhus University, Aarhus, Denmark
[e] Institute of Forensic Medicine, University of Southern Denmark, Odense, Denmark
[f] Department of Forensic Medicine, University of Copenhagen, Copenhagen, Denmark
[g] Department of Forensic Medicine, Aarhus University, Aarhus, Denmark
[h] Department of Biomedicine, Aarhus University, Aarhus, Denmark

ARTICLE INFO

Article history:
Received 9 August 2012
Received in revised form
21 November 2012
Accepted 17 December 2012
Available online 11 January 2013

Keywords:
Suicide
Genetics
Association
Monoamine
Tryptophan
Serotonin

ABSTRACT

Introduction: Strong evidence demonstrates a genetic susceptibility to suicidal behaviour and a relationship between suicide and mental disorders. The aim of this study was to test for association between suicide and five selected genetic variants, which had shown association with suicide in other populations.
Method: We performed a nationwide case-control study on all suicide cases sent for autopsy in Denmark between the years 2000 and 2007. The study comprised 572 cases and 1049 controls and is one of the largest genetic studies in completed suicide to date. The analysed markers were located within the *Serotonin Transporter (SLC6A4), Monoamine Oxidase-A (MAOA)* and the *Tryptophan Hydroxylase I* and *II (TPH1 and TPH2)* genes.
Results: None of the genetic markers within *SLC6A4, MAOA, TPH1* and *TPH2* were significantly associated with completed suicide or suicide method in the basic association tests. Exploratory interaction test showed that the minor allele of rs1800532 in *TPH1* has a protective effect for males younger than 35 years and females older than 50 years, whereas for the oldest male subjects, it tended to be a risk factor. We also observed a significant interaction between age-group and the 5-HTTLPR genotype (with and without rs25531) in *SLC6A4*. The long allele or high expression allele tends to have a protective effect in the middle age-group.
Limitation: We only analysed a limited number of genetic variants.
Conclusion: None of the analysed variants are strong risk factors. To reveal a better understanding of the genes involved in suicide, we suggest future studies should include both genetic and non-genetic factors.

1. Introduction

Suicidal behaviour aggregates in families (Brent et al., 1996; Turecki, 2001), and studies of twins show that monozygotic individuals have a greater concordance for suicide completion and suicide attempts compared to dizygotic individuals (Roy et al., 1991, 1995; Voracek and Loibl, 2007). The heritability is approximately between 30 and 55% (Voracek and Loibl, 2007).

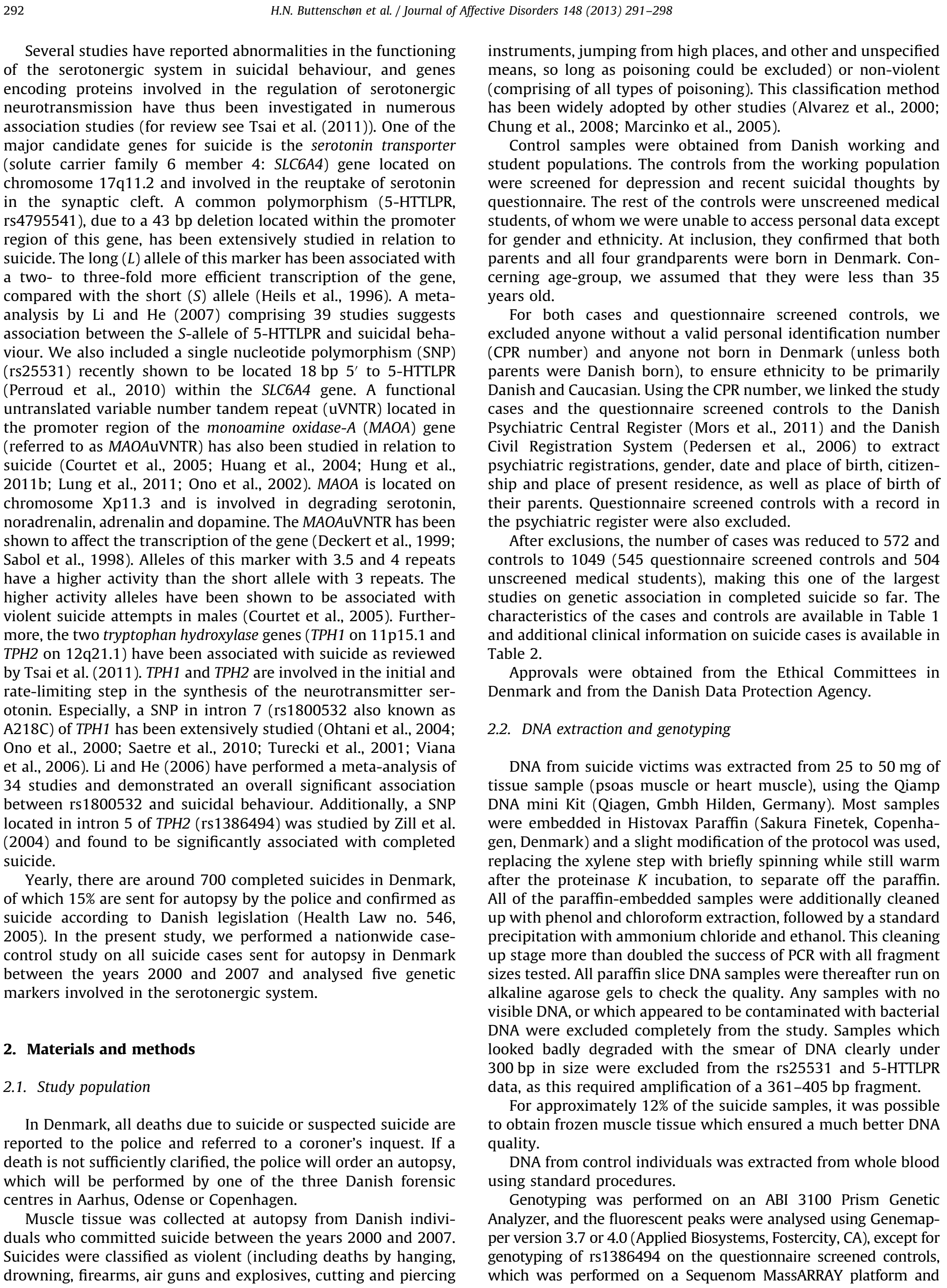
It is well established that psychopathology is an important predictor of suicide completion and that more males than females commit suicide. A meta-analysis comprising 3275 suicides showed, on average, 87% of the subjects who committed suicide had a mental disorder of which affective disorder and any substance disorder were amongst the most common diagnoses (Arsenault-Lapierre et al., 2004). A recent Danish population study, comparing 21,169 suicides over a 17-year period with matched controls, showed that suicide risk is significantly increased for individuals with a hospitalized psychiatric disorder (Qin, 2011).

[*] Corresponding author at: Aarhus University Hospital, Risskov, Centre for Psychiatric Research, Skovagervej 2, DK8240 Risskov, Denmark.
Tel.: +45 7847 1163; fax: +45 7847 1108.
E-mail address: henrbutt@rm.dk (H.N. Buttenschøn).

[1] These authors contributed equally to this work.

0165-0327/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jad.2012.12.011

[33] Journal of Affective Disorders 2013; **148**(2–3): 291–298. © Elsevier. Reused with permission.



Several studies have reported abnormalities in the functioning of the serotonergic system in suicidal behaviour, and genes encoding proteins involved in the regulation of serotonergic neurotransmission have thus been investigated in numerous association studies (for review see Tsai et al. (2011)). One of the major candidate genes for suicide is the *serotonin transporter* (solute carrier family 6 member 4: *SLC6A4*) gene located on chromosome 17q11.2 and involved in the reuptake of serotonin in the synaptic cleft. A common polymorphism (5-HTTLPR, rs4795541), due to a 43 bp deletion located within the promoter region of this gene, has been extensively studied in relation to suicide. The long (*L*) allele of this marker has been associated with a two- to three-fold more efficient transcription of the gene, compared with the short (*S*) allele (Heils et al., 1996). A meta-analysis by Li and He (2007) comprising 39 studies suggests association between the *S*-allele of 5-HTTLPR and suicidal behaviour. We also included a single nucleotide polymorphism (SNP) (rs25531) recently shown to be located 18 bp 5′ to 5-HTTLPR (Perroud et al., 2010) within the *SLC6A4* gene. A functional untranslated variable number tandem repeat (uVNTR) located in the promoter region of the *monoamine oxidase-A* (*MAOA*) gene (referred to as *MAOA*uVNTR) has also been studied in relation to suicide (Courtet et al., 2005; Huang et al., 2004; Hung et al., 2011b; Lung et al., 2011; Ono et al., 2002). *MAOA* is located on chromosome Xp11.3 and is involved in degrading serotonin, noradrenalin, adrenalin and dopamine. The *MAOA*uVNTR has been shown to affect the transcription of the gene (Deckert et al., 1999; Sabol et al., 1998). Alleles of this marker with 3.5 and 4 repeats have a higher activity than the short allele with 3 repeats. The higher activity alleles have been shown to be associated with violent suicide attempts in males (Courtet et al., 2005). Furthermore, the two *tryptophan hydroxylase* genes (*TPH1* on 11p15.1 and *TPH2* on 12q21.1) have been associated with suicide as reviewed by Tsai et al. (2011). *TPH1* and *TPH2* are involved in the initial and rate-limiting step in the synthesis of the neurotransmitter serotonin. Especially, a SNP in intron 7 (rs1800532 also known as A218C) of *TPH1* has been extensively studied (Ohtani et al., 2004; Ono et al., 2000; Saetre et al., 2010; Turecki et al., 2001; Viana et al., 2006). Li and He (2006) have performed a meta-analysis of 34 studies and demonstrated an overall significant association between rs1800532 and suicidal behaviour. Additionally, a SNP located in intron 5 of *TPH2* (rs1386494) was studied by Zill et al. (2004) and found to be significantly associated with completed suicide.

Yearly, there are around 700 completed suicides in Denmark, of which 15% are sent for autopsy by the police and confirmed as suicide according to Danish legislation (Health Law no. 546, 2005). In the present study, we performed a nationwide case-control study on all suicide cases sent for autopsy in Denmark between the years 2000 and 2007 and analysed five genetic markers involved in the serotonergic system.

2. Materials and methods

2.1. Study population

In Denmark, all deaths due to suicide or suspected suicide are reported to the police and referred to a coroner's inquest. If a death is not sufficiently clarified, the police will order an autopsy, which will be performed by one of the three Danish forensic centres in Aarhus, Odense or Copenhagen.

Muscle tissue was collected at autopsy from Danish individuals who committed suicide between the years 2000 and 2007. Suicides were classified as violent (including deaths by hanging, drowning, firearms, air guns and explosives, cutting and piercing instruments, jumping from high places, and other and unspecified means, so long as poisoning could be excluded) or non-violent (comprising of all types of poisoning). This classification method has been widely adopted by other studies (Alvarez et al., 2000; Chung et al., 2008; Marcinko et al., 2005).

Control samples were obtained from Danish working and student populations. The controls from the working population were screened for depression and recent suicidal thoughts by questionnaire. The rest of the controls were unscreened medical students, of whom we were unable to access personal data except for gender and ethnicity. At inclusion, they confirmed that both parents and all four grandparents were born in Denmark. Concerning age-group, we assumed that they were less than 35 years old.

For both cases and questionnaire screened controls, we excluded anyone without a valid personal identification number (CPR number) and anyone not born in Denmark (unless both parents were Danish born), to ensure ethnicity to be primarily Danish and Caucasian. Using the CPR number, we linked the study cases and the questionnaire screened controls to the Danish Psychiatric Central Register (Mors et al., 2011) and the Danish Civil Registration System (Pedersen et al., 2006) to extract psychiatric registrations, gender, date and place of birth, citizenship and place of present residence, as well as place of birth of their parents. Questionnaire screened controls with a record in the psychiatric register were also excluded.

After exclusions, the number of cases was reduced to 572 and controls to 1049 (545 questionnaire screened controls and 504 unscreened medical students), making this one of the largest studies on genetic association in completed suicide so far. The characteristics of the cases and controls are available in Table 1 and additional clinical information on suicide cases is available in Table 2.

Approvals were obtained from the Ethical Committees in Denmark and from the Danish Data Protection Agency.

2.2. DNA extraction and genotyping

DNA from suicide victims was extracted from 25 to 50 mg of tissue sample (psoas muscle or heart muscle), using the Qiamp DNA mini Kit (Qiagen, Gmbh Hilden, Germany). Most samples were embedded in Histovax Paraffin (Sakura Finetek, Copenhagen, Denmark) and a slight modification of the protocol was used, replacing the xylene step with briefly spinning while still warm after the proteinase *K* incubation, to separate off the paraffin. All of the paraffin-embedded samples were additionally cleaned up with phenol and chloroform extraction, followed by a standard precipitation with ammonium chloride and ethanol. This cleaning up stage more than doubled the success of PCR with all fragment sizes tested. All paraffin slice DNA samples were thereafter run on alkaline agarose gels to check the quality. Any samples with no visible DNA, or which appeared to be contaminated with bacterial DNA were excluded completely from the study. Samples which looked badly degraded with the smear of DNA clearly under 300 bp in size were excluded from the rs25531 and 5-HTTLPR data, as this required amplification of a 361–405 bp fragment.

For approximately 12% of the suicide samples, it was possible to obtain frozen muscle tissue which ensured a much better DNA quality.

DNA from control individuals was extracted from whole blood using standard procedures.

Genotyping was performed on an ABI 3100 Prism Genetic Analyzer, and the fluorescent peaks were analysed using Genemapper version 3.7 or 4.0 (Applied Biosystems, Fostercity, CA), except for genotyping of rs1386494 on the questionnaire screened controls, which was performed on a Sequenom MassARRAY platform and

Table 1
Characteristics of controls and suicide cases.

	Controls	Cases
N (proportion)	1049 (0.65)	572 (0.35)
Gender (F/M)	763/286 (0.73/0.27)	209/363 (0.37/0.63)
Age-group[a]	594/218/237 (0.57/0.21/0.23)	128/200/244 (0.22/0.35/0.43)
TPH1: rs1800532[b]	363/483/181 (0.35/0.47/0.18)	182/228/80 (0.37/0.47/0.16)
TPH2: rs1386494[b]	743/268/22 (0.72/0.26/0.021)	396/150/7 (0.72/0.27/0.013)
SLC6A4:		
SS/SL/LL[c]	175/471/391 (0.17/0.45/0.38)	32/138/105 (0.12/0.50/0.38)
6 levels[d]	175/62/5/400/77/307	32/9/2/127/21/81
3 levels[e]	242/477/307 (0.24/0.46/0.30)	43/148/81 (0.16/0.54/0.30)
MAOA: *MAOA*uVNTR		
Alleles[f]	2/634/21/1131/4/7	1/220/6/355/0/10
Male L/H[g]	96/185 (0.34/0.66)	94/160 (0.37/0.63)
Female LL/LH/HH	110/320/329 (0.14/0.42/0.43)	31/65/73 (0.18/0.38/0.43)

[a] Age-group: $<35/35–49/\geq 50$ years.
[b] *TPH1*: CC/CA/AA; *TPH2*: GG/GA/AA.
[c] 5-HTTLPR alone.
[d] 5-HTTLPR combined with rs25531: $SS/SL_G/L_GL_G/SL_A/L_GL_A/L_AL_A$.
[e] $(SS+SL_G+L_GL_G)/(SL_A+L_GL_A)/L_AL_A$: low/medium/high expression.
[f] Alleles of *MAOA*uVNTR corresponding to 2/3/3.5/4/4.5/5 repeats: 187/217/235/247/265/277.
[g] Alleles with low expression, L: 187 and 217; and high, H: 235, 247, 265 and 277.

Table 2
Characteristics of 572 suicide cases and distribution by sex.

Characteristics	**Number (proportion)**	**Female (proportion)**	**Male (proportion)**
Method of suicide (non-violent/violent)	238/334 (0.42/0.58)	111/98 (0.53/0.47)	127/236 (0.35/0.65)
Psychiatric hospital contact (no contact/contact)	247/325 (0.43/0.57)	67/142 (0.32/0.68)	180/183 (0.50/0.50)
Diagnosis of last psychiatric contact (codes in ICD-10)			
Schizophrenia spectrum disorder (F21-F29, F600, F601)	57 (0.10)	25 (0.12)	32 (0.09)
Affective disorder (F30-F39)	59 (0.10)	32 (0.15)	27 (0.07)
Substance dependence (F10-F19)	79 (0.14)	19 (0.09)	60 (0.17)
Reaction to stress/adjustment disorder (F43)	50 (0.09)	30 (0.14)	20 (0.06)
Other psychiatric disorder[a]	80 (0.14)	36 (0.17)	44 (0.12)
Time since last psychiatric contact			
Within 1 year	183 (0.32)	82 (0.39)	101 (0.28)
Within 2–3 years	58 (0.10)	25 (0.12)	33 (0.09)
3 Years ago	84 (0.15)	35 (0.17)	49 (0.14)

[a] Any psychiatric diagnoses not specified.

analysed using the massARRAY Typer 4 software (Sequenom, Inc., San Diego, CA). All genotypes were checked independently by two experienced investigators to reduce the risk of genotyping errors.

The forward primer used for genotyping the 5-HTTLPR and rs25531 markers was modified using a FAM fluoroscein. PCR products were directly used for the 5-HTTLPR genotyping to reveal a long ($L=405$ bp) or short ($S=361$ bp) amplicon size. Five μl of this PCR product were digested for one hour at 37 °C with 0.6 units of the MspI restriction enzyme (New England Biolabs, Ipswich, MA) to reveal the rs25531 SNP (G/A), leading to observed visible fragments according to the fluorescent attached forward primer: $L_A=340$ bp, $L_G=166$ bp and $S=297$ bp. The A to G substitution of this SNP has been shown to modulate the effect of 5-HTTLPR on transcriptional efficacy. The G-allele of this SNP causes the *L*-allele of 5-HTTLPR to behave like the *S*-allele during transcription (Hu et al. ,2006; Kraft et al., 2005).

Using the convention described by Parsey et al. (2006), the combined genotypes of 5-HTTLPR and rs25531 were further reclassified according to functional activity: $SS+SL_G+L_GL_G$ (low expression), $SL_A+L_GL_A$ (medium expression) and L_AL_A (high expression).

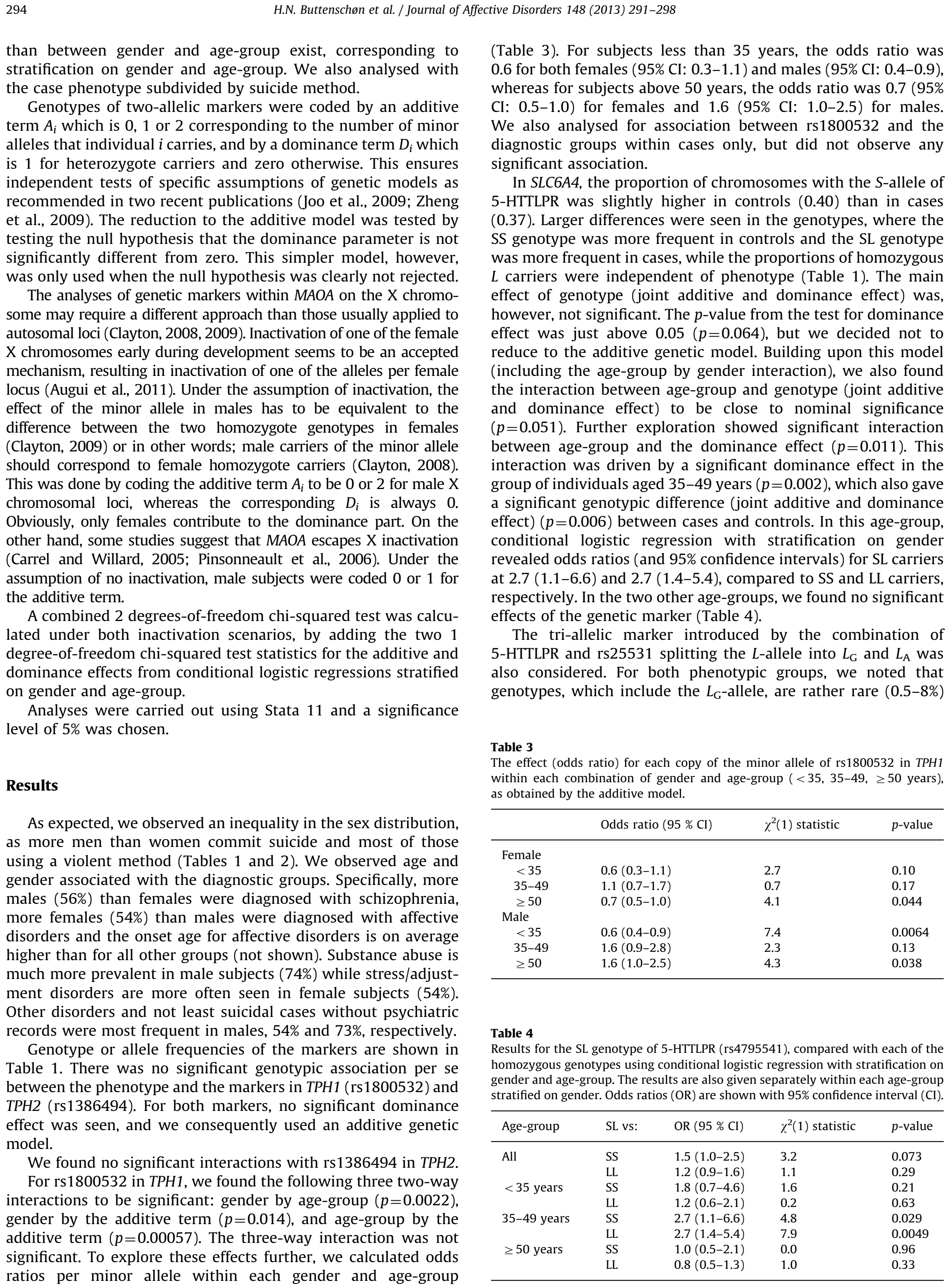
PCR using a FAM fluorescence-labelled primer was used for genotyping *MAOA*uVNTR (Furlong et al., 1999). We also reclassified the *MAOA*uVNTR according to functional activity (Deckert et al., 1999): 2 and 3 repeats (low expression); 3.5, 4, 4.5 and 5 repeats (high expression). The two SNPs, rs1800532 and rs1386494, were genotyped according to the SNaPshot protocol (Applied Biosystems, Fostercity, CA) or the iPLEX Gold reaction protocol (Sequenom, Inc., San Diego, USA).

Additional method conditions and primer sequences are available on request.

Statistics

Association between genetic markers and the case/control phenotype was investigated using logistic regression. In order not to have age as a continuous variable, we grouped age in three groups: <35, 35–49, ≥ 50 years.

Gender, age-group and their interaction were associated with phenotype by sampling (see Table 1), and these factors were therefore included as covariates in the regressions by default. The interaction analyses were not planned a priori. Under certain conditions efficiency would be gained by excluding these factors even when heavily associated with phenotype (Clayton, 2008). We retained them, as we believe that they might be proxies for other unobserved factors, and their interaction was tested with genotype. Results are presented from the relevant conditional logistic regression model, when no other significant interaction



than between gender and age-group exist, corresponding to stratification on gender and age-group. We also analysed with the case phenotype subdivided by suicide method.

Genotypes of two-allelic markers were coded by an additive term A_i which is 0, 1 or 2 corresponding to the number of minor alleles that individual *i* carries, and by a dominance term D_i which is 1 for heterozygote carriers and zero otherwise. This ensures independent tests of specific assumptions of genetic models as recommended in two recent publications (Joo et al., 2009; Zheng et al., 2009). The reduction to the additive model was tested by testing the null hypothesis that the dominance parameter is not significantly different from zero. This simpler model, however, was only used when the null hypothesis was clearly not rejected.

The analyses of genetic markers within *MAOA* on the X chromosome may require a different approach than those usually applied to autosomal loci (Clayton, 2008, 2009). Inactivation of one of the female X chromosomes early during development seems to be an accepted mechanism, resulting in inactivation of one of the alleles per female locus (Augui et al., 2011). Under the assumption of inactivation, the effect of the minor allele in males has to be equivalent to the difference between the two homozygote genotypes in females (Clayton, 2009) or in other words; male carriers of the minor allele should correspond to female homozygote carriers (Clayton, 2008). This was done by coding the additive term A_i to be 0 or 2 for male X chromosomal loci, whereas the corresponding D_i is always 0. Obviously, only females contribute to the dominance part. On the other hand, some studies suggest that *MAOA* escapes X inactivation (Carrel and Willard, 2005; Pinsonneault et al., 2006). Under the assumption of no inactivation, male subjects were coded 0 or 1 for the additive term.

A combined 2 degrees-of-freedom chi-squared test was calculated under both inactivation scenarios, by adding the two 1 degree-of-freedom chi-squared test statistics for the additive and dominance effects from conditional logistic regressions stratified on gender and age-group.

Analyses were carried out using Stata 11 and a significance level of 5% was chosen.

Results

As expected, we observed an inequality in the sex distribution, as more men than women commit suicide and most of those using a violent method (Tables 1 and 2). We observed age and gender associated with the diagnostic groups. Specifically, more males (56%) than females were diagnosed with schizophrenia, more females (54%) than males were diagnosed with affective disorders and the onset age for affective disorders is on average higher than for all other groups (not shown). Substance abuse is much more prevalent in male subjects (74%) while stress/adjustment disorders are more often seen in female subjects (54%). Other disorders and not least suicidal cases without psychiatric records were most frequent in males, 54% and 73%, respectively.

Genotype or allele frequencies of the markers are shown in Table 1. There was no significant genotypic association per se between the phenotype and the markers in *TPH1* (rs1800532) and *TPH2* (rs1386494). For both markers, no significant dominance effect was seen, and we consequently used an additive genetic model.

We found no significant interactions with rs1386494 in *TPH2*.

For rs1800532 in *TPH1*, we found the following three two-way interactions to be significant: gender by age-group ($p=0.0022$), gender by the additive term ($p=0.014$), and age-group by the additive term ($p=0.00057$). The three-way interaction was not significant. To explore these effects further, we calculated odds ratios per minor allele within each gender and age-group (Table 3). For subjects less than 35 years, the odds ratio was 0.6 for both females (95% CI: 0.3–1.1) and males (95% CI: 0.4–0.9), whereas for subjects above 50 years, the odds ratio was 0.7 (95% CI: 0.5–1.0) for females and 1.6 (95% CI: 1.0–2.5) for males. We also analysed for association between rs1800532 and the diagnostic groups within cases only, but did not observe any significant association.

In *SLC6A4*, the proportion of chromosomes with the *S*-allele of 5-HTTLPR was slightly higher in controls (0.40) than in cases (0.37). Larger differences were seen in the genotypes, where the SS genotype was more frequent in controls and the SL genotype was more frequent in cases, while the proportions of homozygous *L* carriers were independent of phenotype (Table 1). The main effect of genotype (joint additive and dominance effect) was, however, not significant. The *p*-value from the test for dominance effect was just above 0.05 ($p=0.064$), but we decided not to reduce to the additive genetic model. Building upon this model (including the age-group by gender interaction), we also found the interaction between age-group and genotype (joint additive and dominance effect) to be close to nominal significance ($p=0.051$). Further exploration showed significant interaction between age-group and the dominance effect ($p=0.011$). This interaction was driven by a significant dominance effect in the group of individuals aged 35–49 years ($p=0.002$), which also gave a significant genotypic difference (joint additive and dominance effect) ($p=0.006$) between cases and controls. In this age-group, conditional logistic regression with stratification on gender revealed odds ratios (and 95% confidence intervals) for SL carriers at 2.7 (1.1–6.6) and 2.7 (1.4–5.4), compared to SS and LL carriers, respectively. In the two other age-groups, we found no significant effects of the genetic marker (Table 4).

The tri-allelic marker introduced by the combination of 5-HTTLPR and rs25531 splitting the *L*-allele into L_G and L_A was also considered. For both phenotypic groups, we noted that genotypes, which include the L_G-allele, are rather rare (0.5–8%)

Table 3
The effect (odds ratio) for each copy of the minor allele of rs1800532 in *TPH1* within each combination of gender and age-group (<35, 35–49, ≥ 50 years), as obtained by the additive model.

	Odds ratio (95 % CI)	$\chi^2(1)$ statistic	*p*-value
Female			
<35	0.6 (0.3–1.1)	2.7	0.10
35–49	1.1 (0.7–1.7)	0.7	0.17
≥ 50	0.7 (0.5–1.0)	4.1	0.044
Male			
<35	0.6 (0.4–0.9)	7.4	0.0064
35–49	1.6 (0.9–2.8)	2.3	0.13
≥ 50	1.6 (1.0–2.5)	4.3	0.038

Table 4
Results for the SL genotype of 5-HTTLPR (rs4795541), compared with each of the homozygous genotypes using conditional logistic regression with stratification on gender and age-group. The results are also given separately within each age-group stratified on gender. Odds ratios (OR) are shown with 95% confidence interval (CI).

Age-group	SL vs:	OR (95 % CI)	$\chi^2(1)$ statistic	*p*-value
All	SS	1.5 (1.0–2.5)	3.2	0.073
	LL	1.2 (0.9–1.6)	1.1	0.29
<35 years	SS	1.8 (0.7–4.6)	1.6	0.21
	LL	1.2 (0.6–2.1)	0.2	0.63
35–49 years	SS	2.7 (1.1–6.6)	4.8	0.029
	LL	2.7 (1.4–5.4)	7.9	0.0049
≥ 50 years	SS	1.0 (0.5–2.1)	0.0	0.96
	LL	0.8 (0.5–1.3)	1.0	0.33

(Table 1). Due to the rarity of the L_G-allele, and by convention from previous publications, we applied a reduced model dictated by the expected equivalent functional behaviour of S and L_G, i.e. by collapsing these two alleles ($S+L_G$). A likelihood ratio test did not reject this reduction ($p=0.90$). In this model, a clear significant dominance effect was observed ($p=0.0065$), and we therefore used the model with the joint additive and dominance effect. As with 5-HTTLPR, there was a tendency of interaction between age-group and genotype ($p=0.077$), and a nominally significant interaction between age-group and the dominance effect ($p=0.049$). Table 5 shows the results from comparing the heterozygous genotype group with each of the homozygous types, using a conditional logistic regression stratifying on both age-group and gender, and for the three age-groups separately (stratifying on gender). We note that, overall; the significant dominance effect is depicted by the fact that the heterozygous group (medium gene expression) raises the risk more than does the homozygous genotype of the L_A-allele ($L_AL_{A,}$ high gene expression). This effect seems to be more pronounced for younger subjects (<35 years), whereas among the oldest individuals (>50 years), the risk effect of the L_A-allele looks more like a clear additive genetic effect, although no differences are statistically significant within this group. Interestingly, in the middle age-group (35–49 years), the L_A-allele tends to have a protective effect but still with a clear dominance effect.

Genotyping *MAOA*uVNTR, we identified a new allele corresponding to a 4.5 repeat of the repeated sequence. The new allele was observed in four control individuals from the group of unscreened medical students. In order to ensure this new allele was not an artefact, the genotyping were repeated independently several times by different investigators. We included the new 4.5 allele into the group of high expression alleles (Table 1). Differences between cases and controls were modest, and the combined 2 degrees-of-freedom test gave a p-value of 0.19 and 0.15 in the model assuming X-inactivation, and no inactivation, respectively. We found no significant interactions, neither between age-group and genotypes, nor between age-group and any of the additive and dominance parameters. Using conditional logistic regression, with stratification on age-group, the additive term was not significant in the separate samples of male and female individuals. In the genotypic model for females, the dominance term was just above the border of significance ($p=0.067$).

With respect to suicide method, we observed no differences between the two case groups for any of the analysed genetic markers.

Table 5
Results for the (SL_A and L_GL_A) genotypic marker combination in *SLC6A4* with alleles grouped by function. $SS+SL_G+L_GL_G$ correspond to low gene expression, $SL_A+L_GL_A$ to medium gene expression and L_AL_A to high gene expression. The heterozygote genotype is compared with each of the homozygous categories, using conditional logistic regression with stratification on gender and age-group. The results are also given separately within each age-group stratified on gender. Odds ratios (OR) are shown with 95% confidence interval (CI).

Age-group	$SL_A+L_GL_A$ (medium) vs:	OR (95% CI)	$\chi^2(1)$ statistic	p-value
All	$SS+SL_G+L_GL_G$ (low)	1.7 (1.1–2.6)	6.4	0.011
	L_AL_A (high)	1.4 (1.0–1.9)	3.1	0.079
<35 Years	$SS+SL_G+L_GL_G$ (low)	2.8 (1.2–6.5)	5.5	0.019
	L_AL_A (high)	1.9 (1.0–3.6)	3.5	0.063
35–49 Years	$SS+SL_G+L_GL_G$ (low)	1.7 (0.8–3.7)	2.0	0.16
	L_AL_A (high)	2.4 (1.2–4.8)	5.5	0.019
≥ 50 Years	$SS+SL_G+L_GL_G$ (low)	1.4 (0.7–2.6)	0.8	0.37
	L_AL_A (high)	0.8 (0.5–1.4)	0.6	0.45

Discussion

We have performed a large nationwide association study on suicides sent for autopsy in Denmark and analysed five genetic variants located within the *SLC6A4,* the *MAOA* and the *TPH1* and *TPH2* genes, respectively.

Numerous studies have tested for association between rs1800532 (A218C) in *TPH1* and suicide, but produced contrary results. In a meta-analysis from 2006, an increased proportion of the A-allele was found in cases with suicidal behaviour compared with control individuals (Li and He, 2006). But several studies on completed suicides did not find a positive association to this marker (Bennett et al., 2000; Ohtani et al., 2004; Ono et al., 2000; Stefulj et al., 2005). In accordance with these studies, we did not observe any significant association between rs1800532 in *TPH1* and completed suicide. Further analyses, however, showed that the effect of carrying the A-allele depends both on gender and age-group, as both two-way interactions with the additive genetic term were significant. Table 3 shows how this interaction manifests with a clearly significant protective effect of the minor A-allele in male subjects younger than 35 years, while the A-allele tends to be a risk factor for male subjects older than 35 years. In contrast to this, a protective effect was observed for females in the oldest age-group.

We could not replicate Zill et al. (2004) finding of association between the G-allele of rs1386494 in *TPH2* and completed suicide. The very low frequency of the homozygous minor allele genotype unfortunately limited the possibilities for doing reasonable analyses with more than a few parameters.

The 5-HTTLPR genetic marker located within the *SLC6A4* gene has previously been extensively studied and has shown conflicting results. Some studies have reported association between the *S*-allele and suicide (Courtet et al., 2001; Neves et al., 2010; Segal et al., 2006), while others find no statistical significant difference (Coventry et al., 2010; Geijer et al., 2000; Rujescu et al., 2001; Shen et al., 2004). A meta-analysis from 2007 shows support for association between 5-HTTLPR and suicidal behaviour, but in the subgroup analyses of studies, including only suicide completers compared to healthy control individuals, no association was observed (Li and He, 2007). Likewise, we did not find evidence of association between 5-HTTLPR and suicide in the basic association tests. We could not replicate previous findings suggesting association between violent completed suicide and the *S*-allele of 5-HTTLPR, as recently reviewed by Gonda et al. (2011). A few studies have investigated the 5-HTTLPR in combination with rs25531 (De Luca et al., 2006b; Hung et al., 2011a; Segal et al., 2009). A recent study on Chinese suicide attempters showed a significant association to this tri-allelic marker (Hung et al., 2011a). This is in contrast to our study and other studies using subjects of Caucasian origin (De Luca et al., 2006b; Segal et al., 2009).

The exploratory interaction analyses of the 5-HTTLPR marker in the present study showed that the effect of being heterozygous depends on age-group. An elevated suicide risk was observed for heterozygous individuals between 35 and 49 years compared to homozygous individuals. Further exploration of this revealed that this effect was more pronounced in males than in females (results not shown). The interaction analyses of the tri-allelic marker in the serotonin transporter showed a statistically significant protective effect of the low expression genotypes for individuals below 35 years, and a statistically significant protective effect of the high expression genotype for individuals between 35 and 49 years. These exploratory interaction analyses are not obvious to interpret, and may reflect some underlying unobserved factors. As far as we know, similar analyses have not been performed by others.

We found no association between *MAOA*uVNTR and suicide. A few previous studies have suggested an association between the polymorphism and suicidal behaviour (Ho et al., 2000; Lung et al., 2011), but our results are consistent with most studies, and in addition to a recent meta-analysis, showing no association (De Luca et al., 2005, 2006a; Huang et al., 2004; Hung et al., 2011b; Ono et al., 2002). To our knowledge, only one previous study included completed suicide subjects, and found no association (Ono et al., 2002). A study by Courtet et al. (2005) showed the frequency of the high activity alleles (3.5 and 4 repeats, respectively) to be higher in men who attempted violent suicide compared to men who used non-violent means. We were not able to replicate that the method of suicide was influenced by the genotype, however.

Recently genome-wide association studies on attempted suicide or suicidal thought have been published but did not shown support for *SLC6A4*, *MAOA*, *TPH1* or *TPH2* (Perlis et al., 2010; Schosser et al., 2011; Willour et al., 2012). Judy et al. (2012) genotyped 174 tag and coding SNPs located in genes within the serotonergic pathway on more than 500 individual with a history of attempted suicide and more than 500 healthy control individuals. This study included rs1800532 within *TPH1* and rs1386494 within *TPH2* as in our study, but did not include the 5-HTTLPR or the *MAOA*uVNTR polymorphism. In summary, the study showed a few nominal significant association signals, none of which survived correction for multiple testing. Neither rs1800532 nor rs1386494 was amongst these nominal associated SNPs.

In agreement with the literature, the suicides from our study included more males than females (Heuveline and Slap, 2002). More females than males had a history of contact with a psychiatric hospital, however. In total 57% of all suicide cases had a history of contact and most within one year since last contact. This is in agreement with the large population based study by Qin (2011) comparing more than 21,000 suicides in Denmark, where 37% of male and 57% of female suicide individuals had a recorded history of psychiatric hospitalization. In Denmark, all residents have equal access to psychiatric hospitals, and the treatment is free of charge, ensuring that all psychiatric admissions are represented in the register. A larger percentage of suicides with a psychiatric disorder were reported in the meta-analysis by Arsenault-Lapierre et al. (2004). This might be explained by the fact that diagnoses of the suicide completers included in the meta-analysis were reconstructed based on interviews with the informants or on reviews of official records whereas we only included information on diagnoses from the psychiatric register. In agreement with other studies (Qin, 2011; Arsenault-Lapierre et al., 2004) we found affective disorders more common among female suicide completers whereas substance-related disorders were more common among male suicide completers.

Our study, however, should be viewed in the light of several limitations. First, amplification of DNA that has been extracted from paraffin blocks from suicide victim samples gave some technical problems for the longer fragments especially for 5-HTTLPR. We had no problems of this sort with the frozen tissue, unless the tissue was very badly degraded before freezing. These problems led to exclusions of cases and genotypes of the 5-HTTLPR genetic marker. Despite this, the present genetic study on 5-HTTLPR is a large association study on completed suicide. Overall, 10% of the suicide cases were excluded completely from the study due to degraded tissue. Second, the study was designed to test for association with a limited number of genetic variants. Thus we did not capture most of the genetic variants within the selected genes nor did we test for association with rare variants. Third, approximately half of the controls were unscreened medical students, of whom we were unable to access personal data except for gender and ethnicity. These controls were included in the study to ensure more even numbers of male cases and controls. The distribution on gender is still different between cases and controls, but this can be a caveat of using standard control sets, e.g. the common set of controls used for several diseases in the Wellcome Trust Case Control Consortium (WTCCC) (2007). Fourth, even though this is one of the largest studies on completed suicide the sample size is too small when considering, e.g., investigations in subgroups determined by psychiatric diagnosis.

In conclusion, this is one of the largest studies on completed suicide. We investigated five genetic markers located within four genes involved in the serotonergic system for association to suicide, but did not find evidence of association in the basic association test. We performed exploratory interaction analyses, however, and observed significant two-way interactions for rs1800532 in *TPH1*, and for 5-HTTLPR (and rs25531) in the *Serotonin Transporter*. The minor allele of rs1800532 in *TPH1* showed a protective effect against committing suicide for different age-groups depending on gender. Having a heterozygous genotype of the 5-HTTLPR marker, or a heterozygous genotype of the tri-allelic marker (5-HTTLPR in combination with rs25531), was also a risk factor for suicide completion in individuals between 35 and 49 years.

Our findings suggest that none of the five genetic variants included in the present study are strong risk factors. However, the additional interaction analyses indicate the importance of age and gender. Future large studies on suicide including non-genetic factors are warranted. Suicidal behaviour is a complex phenotype, and a better understanding of the genes involved in suicide and their interactions may help in describing risk factors.

Role of funding source

The PRISME project was supported by 'Arbejdsmiljøforskningsfonden', 'TrygFonden' and The Lundbeck Foundation in Denmark. The present study was supported by 'Jascha Fonden' and 'Forskningsfonden til støtte af psykiatrisk forskning i Region Midtjylland' in Denmark.

Conflict of interest

All authors reported no biomedical financial interests or other potential conflicts of interest.

Acknowledgements

We acknowledge the PRISME group for collection of control DNA and documentation, and Helle Nygaard Buch, Margit Palmer, Agata El Daoud, Eva Olesen and Jytte Jakobsen for technical assistance. The authors acknowledge Maja Strand and Gitte Murmand who assisted in proof-reading of the manuscript.

References

Alvarez, J.C., Cremniter, D., Gluck, N., Quintin, P., Leboyer, M., Berlin, I., Therond, P., Spreux-Varoquaux, O., 2000. Low serum cholesterol in violent but not in non-violent suicide attempters. Psychiatric Research 95 (2), 103–108.

Arsenault-Lapierre, G., Kim, C., Turecki, G., 2004. Psychiatric diagnoses in 3275 suicides: a meta-analysis. BMC. Psychiatry 4, 37.

Augui, S., Nora, E.P., Heard, E., 2011. Regulation of X-chromosome inactivation by the X-inactivation centre. Nature Reviews Genetic 12 (6), 429–442.

Bennett, P.J., McMahon, W.M., Watabe, J., Achilles, J., Bacon, M., Coon, H., Grey, T., Keller, T., Tate, D., Tcaciuc, I., Workman, J., Gray, D., 2000. Tryptophan hydroxylase polymorphisms in suicide victims. Psychiatric Genetics 10 (1), 13–17.

Brent, D.A., Bridge, J., Johnson, B.A., Connolly, J., 1996. Suicidal behavior runs in families. A controlled family study of adolescent suicide victims. Archives of General Psychiatry 53 (12), 1145–1152.

Carrel, L., Willard, H.F., 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434 (7031), 400–404.

Chung, K.H., Lee, H.C., Kao, S., Lin, H.C., 2008. Urbanicity and methods of suicide: a nationwide population-based study. Journal of Urban Health 85 (1), 136–145.

Clayton, D., 2008. Testing for association on the X chromosome. Biostatistics 9 (4), 593–600.

Clayton, D.G., 2009. Sex chromosomes and genetic association studies. Genome Medicine 1 (11), 110.

Courtet, P., Baud, P., Abbar, M., Boulenger, J.P., Castelnau, D., Mouthon, D., Malafosse, A., Buresi, C., 2001. Association between violent suicidal behavior and the low activity allele of the serotonin transporter gene. Molecular Psychiatry 6 (3), 338–341.

Courtet, P., Jollant, F., Buresi, C., Castelnau, D., Mouthon, D., Malafosse, A., 2005. The monoamine oxidase A gene may influence the means used in suicide attempts. Psychiatric Genetics 15 (3), 189–193.

Coventry, W.L., James, M.R., Eaves, L.J., Gordon, S.D., Gillespie, N.A., Ryan, L., Heath, A.C., Montgomery, G.W., Martin, N.G., Wray, N.R., 2010. Do 5HTTLPR and stress interact in risk for depression and suicidality? Item response analyses of a large sample. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 153B (3), 757–765.

De Luca, V., Tharmalingam, S., Muller, D.J., Wong, G., de, B.A., Kennedy, J.L., 2006a. Gene–gene interaction between MAOA and COMT in suicidal behavior: analysis in schizophrenia. Brain Research 1097 (1), 26–30.

De Luca, V., Tharmalingam, S., Sicard, T., Kennedy, J.L., 2005. Gene–gene interaction between MAOA and COMT in suicidal behavior. Neuroscience Letters 383 (1-2), 151–154.

De Luca, V., Zai, G., Tharmalingam, S., de, B.A., Wong, G., Kennedy, J.L., 2006b. Association study between the novel functional polymorphism of the serotonin transporter gene and suicidal behaviour in schizophrenia. European Neuropsychopharmacology 16 (4), 268–271.

Deckert, J., Catalano, M., Syagailo, Y.V., Bosi, M., Okladnova, O., Di, B.D., Nothen, M.M., Maffei, P., Franke, P., Fritze, J., Maier, W., Propping, P., Beckmann, H., Bellodi, L., Lesch, K.P., 1999. Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder. Human Molecular Genetics 8 (4), 621–624.

Furlong, R.A., Ho, L., Rubinsztein, J.S., Walsh, C., Paykel, E.S., Rubinsztein, D.C., 1999. Analysis of the monoamine oxidase A (MAOA) gene in bipolar affective disorder by association studies, meta-analyses, and sequencing of the promoter. American Journal of Medical Genetics 88 (4), 398–406.

Geijer, T., Frisch, A., Persson, M.L., Wasserman, D., Rockah, R., Michaelovsky, E., Apter, A., Jonsson, E.G., Nothen, M.M., Weizman, A., 2000. Search for association between suicide attempt and serotonergic polymorphisms. Psychiatric Genetics 10 (1), 19–26.

Gonda, X., Fountoulakis, K.N., Harro, J., Pompili, M., Akiskal, H.S., Bagdy, G., Rihmer, Z., 2011. The possible contributory role of the *S* allele of 5-HTTLPR in the emergence of suicidality. Journal of Psychopharmacology 25 (7), 857–866.

Health Law no. 546 (2005) The Danish Government section XIII (§179 and § 184).

Heils, A., Teufel, A., Petri, S., Stober, G., Riederer, P., Bengel, D., Lesch, K.P., 1996. Allelic variation of human serotonin transporter gene expression. Journal of Neurochemistry 66 (6), 2621–2624.

Heuveline, P., Slap, G.B., 2002. Adolescent and young adult mortality by cause: age, gender, and country, 1955 to 1994. Journal of Adolescent Health 30 (1), 29–34.

Ho, L.W., Furlong, R.A., Rubinsztein, J.S., Walsh, C., Paykel, E.S., Rubinsztein, D.C., 2000. Genetic associations with clinical characteristics in bipolar affective disorder and recurrent unipolar depressive disorder. American Journal of Medical Genetics 96 (1), 36–42.

Hu, X.Z., Lipsky, R.H., Zhu, G., Akhtar, L.A., Taubman, J., Greenberg, B.D., Xu, K., Arnold, P.D., Richter, M.A., Kennedy, J.L., Murphy, D.L., Goldman, D., 2006. Serotonin transporter promoter gain-of-function genotypes are linked to obsessive–compulsive disorder. American Journal of Human Genetics 78 (5), 815–826.

Huang, Y.Y., Cate, S.P., Battistuzzi, C., Oquendo, M.A., Brent, D., Mann, J.J., 2004. An association between a functional polymorphism in the monoamine oxidase a gene promoter, impulsive traits and early abuse experiences. Neuropsychopharmacology 29 (8), 1498–1505.

Hung, C.F., Lung, F.W., Chen, C.H., O'Nions, E., Hung, T.H., Chong, M.Y., Wu, C.K., Wen, J.K., Lin, P.Y., 2011a. Association between suicide attempt and a tri-allelic functional polymorphism in serotonin transporter gene promoter in Chinese patients with schizophrenia. Neuroscience Letters 504 (3), 242–246.

Hung, C.F., Lung, F.W., Hung, T.H., Chong, M.Y., Wu, C.K., Wen, J.K., Lin, P.Y., 2011b. Monoamine oxidase A gene polymorphism and suicide: an association study and meta-analysis. Journal of Affective Disorders.

Joo, J., Kwak, M., Ahn, K., Zheng, G., 2009. A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. Biometrics 65 (4), 1115–1122.

Judy, J.T., Seifuddin, F., Mahon, P.B., Huo, Y., Goes, F.S., Jancic, D., Schweizer, B., Mondimore, F.M., Mackinnon, D.F., Depaulo Jr., J.R., Gershon, E.S., McMahon, F.J., Cutler, D.J., Zandi, P.P., Potash, J.B., Willour, V.L., 2012. Association study of serotonin pathway genes in attempted suicide. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 159 (1), 112–119.

Kraft, J.B., Slager, S.L., McGrath, P.J., Hamilton, S.P., 2005. Sequence analysis of the serotonin transporter and associations with antidepressant response. Biological Psychiatry 58 (5), 374–381.

Li, D., He, L., 2007. Meta-analysis supports association between serotonin transporter (5-HTT) and suicidal behavior. Molecular Psychiatry 12 (1), 47–54.

Li, D., He, L., 2006. Further clarification of the contribution of the tryptophan hydroxylase (TPH) gene to suicidal behavior using systematic allelic and genotypic meta-analyses. Human Genetics 119 (3), 233–240.

Lung, F.W., Tzeng, D.S., Huang, M.F., Lee, M.B., 2011. Association of the MAOA promoter uVNTR polymorphism with suicide attempts in patients with major depressive disorder. BMC Medical Genetics 12, 74.

Marcinko, D., Martinac, M., Karlovic, D., Filipcic, I., Loncar, C., Pivac, N., Jakovljevic, M., 2005. Are there differences in serum cholesterol and cortisol concentrations between violent and non-violent schizophrenic male suicide attempters? Collegium Antropologicum 29 (1), 153–157.

Mors, O., Perto, G.P., Mortensen, P.B., 2011. The Danish Psychiatric Central Research Register. Scandinavian Journal of Public Health 39 (7 Suppl), 54–57.

Neves, F.S., Malloy-Diniz, L.F., Romano-Silva, M.A., Aguiar, G.C., de Matos, L.O., Correa, H., 2010. Is the serotonin transporter polymorphism (5-HTTLPR) a potential marker for suicidal behavior in bipolar disorder patients? Journal of Affective Disorders 125 (1-3), 98–102.

Ohtani, M., Shindo, S., Yoshioka, N., 2004. Polymorphisms of the tryptophan hydroxylase gene and serotonin 1A receptor gene in suicide victims among Japanese. Tohoku Journal of Experimental Medicine 202 (2), 123–133.

Ono, H., Shirakawa, O., Nishiguchi, N., Nishimura, A., Nushida, H., Ueno, Y., Maeda, K., 2002. No evidence of an association between a functional monoamine oxidase a gene polymorphism and completed suicides. American Journal of Medical Genetics 114 (3), 340–342.

Ono, H., Shirakawa, O., Nishiguchi, N., Nishimura, A., Nushida, H., Ueno, Y., Maeda, K., 2000. Tryptophan hydroxylase gene polymorphisms are not associated with suicide. American Journal of Medical Genetics 96 (6), 861–863.

Parsey, R.V., Hastings, R.S., Oquendo, M.A., Hu, X., Goldman, D., Huang, Y.Y., Simpson, N., Arcement, J., Huang, Y., Ogden, R.T., Van Heertum, R.L., Arango, V., Mann, J.J., 2006. Effect of a triallelic functional polymorphism of the serotonin-transporter-linked promoter region on expression of serotonin transporter in the human brain. American Journal of Psychiatry 163 (1), 48–51.

Pedersen, C.B., Gotzsche, H., Moller, J.O., Mortensen, P.B., 2006. The Danish Civil Registration System. A cohort of eight million persons. Danish Medical Bulletin 2006 (53), 441–449.

Perlis, R.H., Huang, J., Purcell, S., Fava, M., Rush, A.J., Sullivan, P.F., Hamilton, S.P., McMahon, F.J., Schulze, T.G., Potash, J.B., Zandi, P.P., Willour, V.L., Penninx, B.W., Boomsma, D.I., Vogelzangs, N., Middeldorp, C.M., Rietschel, M., Nothen, M., Cichon, S., Gurling, H., Bass, N., McQuillin, A., Hamshere, M., Craddock, N., Sklar, P., Smoller, J.W., 2010. Genome-wide association study of suicide attempts in mood disorder patients. American Journal of Psychiatry 167 (12), 1499–1507.

Perroud, N., Salzmann, A., Saiz, P.A., Baca-Garcia, E., Sarchiapone, M., Garcia-Portilla, M.P., Carli, V., Vaquero-Lorenzo, C., Jaussent, I., Mouthon, D., Vessaz, M., Huguelet, P., Courtet, P., Malafosse, A., 2010. Rare genotype combination of the serotonin transporter gene associated with treatment response in severe personality disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 153B (8), 1494–1497.

Pinsonneault, J.K., Papp, A.C., Sadee, W., 2006. Allelic mRNA expression of X-linked monoamine oxidase a (MAOA) in human brain: dissection of epigenetic and genetic factors. Human Molecular Genetics 15 (17), 2636–2649.

Qin, P., 2011. The impact of psychiatric illness on suicide: differences by diagnosis of disorders and by sex and age of subjects. Journal of Psychiatric Research 45 (11), 1445–1452.

Roy, A., Segal, N.L., Centerwall, B.S., Robinette, C.D., 1991. Suicide in twins. Archives of General Psychiatry 48 (1), 29–32.

Roy, A., Segal, N.L., Sarchiapone, M., 1995. Attempted suicide among living co-twins of twin suicide victims. American Journal of Psychiatry 152 (7), 1075–1076.

Rujescu, D., Giegling, I., Sato, T., Moeller, H.J., 2001. A polymorphism in the promoter of the serotonin transporter gene is not associated with suicidal behavior. Psychiatric Genetics 11 (3), 169–172.

Sabol, S.Z., Hu, S., Hamer, D., 1998. A functional polymorphism in the monoamine oxidase A gene promoter. Human Genetics 103 (3), 273–279.

Saetre, P., Lundmark, P., Hansen, T., Rasmussen, H.B., Djurovic, S., Melle, I., Andreassen, O.A., Werge, T., Agartz, I., Hall, H., Terenius, L., Jonsson, E.G., 2010. The tryptophan hydroxylase 1 (TPH1) gene, schizophrenia susceptibility, and suicidal behavior: a multi-centre case-control study and meta-analysis. American Journal of Medical Genetics B Neuropsychiatric Genetics 153B (2), 387–396.

Schosser, A., Butler, A.W., Ising, M., Perroud, N., Uher, R., Ng, M.Y., Cohen-Woods, S., Craddock, N., Owen, M.J., Korszun, A., Jones, L., Jones, I., Gill, M., Rice, J.P., Maier, W., Mors, O., Rietschel, M., Lucae, S., Binder, E.B., Preisig, M., Perry, J., Tozzi, F., Muglia, P., Aitchison, K.J., Breen, G., Craig, I.W., Farmer, A.E., Muller-Myhsok, B., McGuffin, P., Lewis, C.M., 2011. Genomewide association scan of suicidal thoughts and behaviour in major depression. PLoS One 6 (7), e20690.

Segal, J., Pujol, C., Birck, A., Gus, M.G., Leistner-Segal, S., 2006. Association between suicide attempts in south Brazilian depressed patients with the serotonin transporter polymorphism. Psychiatry Research 143 (2-3), 289–291.

Segal, J., Schenkel, L.C., Oliveira, M.H., Salum, G.A., Bau, C.H., Manfro, G.G., Leistner-Segal, S., 2009. Novel allelic variants in the human serotonin transporter gene linked polymorphism (5-HTTLPR) among depressed patients with suicide attempt. Neuroscience Letters 451 (1), 79–82.

Shen, Y., Li, H., Gu, N., Tan, Z., Tang, J., Fan, J., Li, X., Sun, W., He, L., 2004. Relationship between suicidal behavior of psychotic inpatients and serotonin transporter gene in Han Chinese. Neuroscience Letters 372 (1-2), 94–98.

Stefulj, J., Kubat, M., Balija, M., Skavic, J., Jernej, B., 2005. Variability of the tryptophan hydroxylase gene: study in victims of violent suicide. Psychiatry Research 134 (1), 67–73.

Tsai, S.J., Hong, C.J., Liou, Y.J., 2011. Recent molecular genetic studies and methodological issues in suicide research. Progress in Neuro-psychopharmacology and Biological Psychiatry 35 (4), 809–817.

Turecki, G., 2001. Suicidal behavior: is there a genetic predisposition? Bipolar Disorders 3 (6), 335–349.

Turecki, G., Zhu, Z., Tzenova, J., Lesage, A., Seguin, M., Tousignant, M., Chawky, N., Vanier, C., Lipp, O., Alda, M., Joober, R., Benkelfat, C., Rouleau, G.A., 2001. TPH and suicidal behavior: a study in suicide completers. Molecular Psychiatry 6 (1), 98–102.

Viana, M.M., De Marco, L.A., Boson, W.L., Romano-Silva, M.A., Correa, H., 2006. Investigation of A218C tryptophan hydroxylase polymorphism: association with familial suicide behavior and proband's suicide attempt characteristics. Genes, Brain and Behavior 5 (4), 340–345.

Voracek, M., Loibl, L.M., 2007. Genetics of suicide: a systematic review of twin studies. Wiener klinische Wochenschrift 119 (15-16), 463–475.

Wellcome Trust Case Control Consortium (WTCCC), 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145), 661–678.

Willour, V.L., Seifuddin, F., Mahon, P.B., Jancic, D., Pirooznia, M., Steele, J., Schweizer, B., Goes, F.S., Mondimore, F.M., Mackinnon, D.F., Perlis, R.H., Lee, P.H., Huang, J., Kelsoe, J.R., Shilling, P.D., Rietschel, M., Nothen, M., Cichon, S., Gurling, H., Purcell, S., Smoller, J.W., Craddock, N., Depaulo Jr., J.R., Schulze, T.G., McMahon, F.J., Zandi, P.P., Potash, J.B., 2012. A genome-wide association study of attempted suicide. Molecular Psychiatry 17 (4), 433–444.

Zheng, G., Joo, J., Yang, Y., 2009. Pearson's test, trend test, and MAX are all trend tests with different types of scores. Annals of Human Genetics 73 (2), 133–140.

Zill, P., Buttner, A., Eisenmenger, W., Moller, H.J., Bondy, B., Ackenheil, M., 2004. Single nucleotide polymorphism and haplotype analysis of a novel tryptophan hydroxylase isoform (TPH2) gene in suicide victims. Biological Psychiatry 56 (8), 581–586.

6.5 Paper 5[34]

Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3

Henriette Nørmølle Buttenchøn[a], Leslie Foldager[a,b], Tracey J. Flint[a], Inger Marie L. Olsen[a], Thomas Deleuran[a], Mette Nyegaard[c], Mette M. Hansen[a], Pekka Kallunki[d], Kenneth V. Christensen[d], Douglas H. Blackwood[f], Walter J. Muir[f], Steen E. Straarup[a], Thomas D. Als[a], Merete Nordentoft[e], Anders D. Børglum[a,c] and Ole Mors[a]

Objective **Linkage and association studies of bipolar affective disorder (BAD) point out chromosome 12q24 as a region of interest.**

Methods **To investigate this region further, we conducted an association study of 22 DNA markers within a 1.14 Mb region in a Danish sample of 166 patients with BAD and 311 control individuals. Two-hundred and four Danish patients with schizophrenia were also included in the study.**

Results **We observed highly significant allelic and genotypic association between BAD and two highly correlated markers. The risk allele of both markers considered separately conferred an odds ratio of 2 to an individual carrying one risk allele and an odds ratio of 4 for individuals carrying both risk alleles assuming an additive genetic model. These findings were supported by the haplotype analysis. In addition, we obtained a replication of four markers associated with BAD in an earlier UK study.**

The most significantly associated marker was also analyzed in a Scottish case–control sample and was earlier associated with BAD in the UK cohort. The association of that particular marker was strongly associated with BAD in a meta-analysis of the Danish, Scottish and UK sample ($P=0.0003$).

The chromosome region confined by our most distant markers is gene-poor and harbours only a few predicted genes. This study implicates the Slynar locus. We confirmed one annotated Slynar transcript and identified a novel transcript in human brain cDNA.

Conclusion **This study confirms 12q24.3 as a region of functional importance in the pathogenesis of BAD and highlights the importance of focused genotyping. *Psychiatr Genet* 20:93–101 © 2010 Wolters Kluwer Health | Lippincott Williams & Wilkins.**

Psychiatric Genetics 2010, **20**:93–101

Keywords: 12q24.3, association, bipolar affective disorder, schizophrenia, Slynar

[a]Centre for Psychiatric Research, Aarhus University Hospital, Risskov, [b]Bioinformatics Research Centre, Aarhus University, Aarhus, [c]Institute of Human Genetics, Aarhus University, Aarhus, [d]H. Lundbeck A/S, Valby, [e]Psychiatric Centre Bispebjerg, University of Copenhagen, Copenhagen, Denmark and [f]Department of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK

Correspondence to Henriette Nørmølle Buttenchøn, PhD, Centre for Psychiatric Research, Aarhus University Hospital, Skovagervej 2, Risskov 8240, Denmark Tel: +45 77893575; fax: +45 77893599; e-mail: henrbutt@rm.dk

Henriette Nørmølle Buttenchøn and Leslie Foldager contributed equally to this study

Received 24 June 2009 **Revised** 2 December 2009
Accepted 7 December 2009

Introduction

Bipolar affective disorder (BAD) is a complex disorder, with a unelectable genetic component, but the exact genetic background remains unidentified (Kato, 2007). We have earlier, in two separate populations (Danish and Faroese), identified a region on chromosome 12q24 linked and associated with BAD (Ewald *et al.*, 1998; Degn *et al.*, 2001; Ewald *et al.*, 2002). At microsatellite marker, D12S1639, a genome-wide significant logarithm of odds score (lod score) was obtained (Ewald *et al.*, 2002). This chromosomal region has been further identified by others as a susceptibility locus for BAD (Morissette *et al.*, 1999; Jones *et al.*, 2002; Curtis *et al.*, 2003; Lyons-Warren *et al.*, 2005; Shink *et al.*, 2005a, 2005b) and has also been implicated in other psychiatric disorders such as schizophrenia (SZ) (Shinkai *et al.*, 2002; Reif *et al.*, 2006) and Alzheimer's disease (Zubenko *et al.*, 1999). This suggests that 12q24 may harbour one or more genes important for a number of mental disorders.

Kalsi *et al.* (2006) performed fine mapping of the 12q24 linkage findings in two cohorts from England and Denmark, and found a significant association with markers surrounding the microsatellite marker, D12S307. The signal was located within the Slynar locus in a 300 kb candidate region. The results, however, only showed region-wise replication between the Danish and English samples. Eighty-one BAD cases and 120 controls from this

Supplementary data are available directly from the authors.

0955-8829 © 2010 Wolters Kluwer Health | Lippincott Williams & Wilkins DOI: 10.1097/YPG.0b013e32833a2066

[34]Psychiatric Genetics 2010; **20**(3): 93–101. © Wolters Kluwer Health Lippincott Williams & Wilkins. Reused with permission.

study were earlier included in the Danish sample in Kalsi *et al.* (2006), where they were genotyped for microsatellites only.

The purpose of this study was to perform a thorough reexamination of the Slynar locus on 12q24.3 in a Danish sample of patients with BAD and SZ and control individuals to (i) perform a replication (ii) fine map the Slynar locus by selection of additional single nucleotide polymorphisms (SNPs) based on a functional approach and (iii) investigate whether the locus is a common susceptibility locus for BAD and SZ.

For further replication, the most significantly associated marker was genotyped and analyzed in 162 Scottish patients with BAD and 200 Scottish controls. We also set out to identify the most abundant Slynar transcripts both in human brain and other tissues, and to identify possible novel transcripts.

Methods

Assessment of cases and controls

The Danish case–control sample consisted of 166 patients with BAD, 204 patients with SZ and 311 ethnically matched controls.

Cases were interviewed with the semistructured diagnostic interview Schedules for Clinical Assessment in Neuropsychiatry (version 2.1) (World Health Organization, 1998) and final best-estimate life-time diagnoses were achieved by consensus of two experienced psychiatrists.

The patients with BAD fulfilled the International Classification of Diseases, Tenth Revision Diagnostic Criteria for Research (World Health Organization, 1993) for BAD and the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (American Psychiatric Association, 1994) criteria for bipolar 1 disorder.

The individuals with SZ fulfilled the International Classification of Diseases, Tenth Revision Diagnostic Criteria for Research (World Health Organization, 1993) and the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (American Psychiatric Association, 1994) criteria for SZ.

Control individuals were unscreened for psychiatric disorders. Both cases and controls were of Danish Caucasian descent three generations back.

The Scottish case–control sample consisted of 162 patients with BAD and 200 ethnically matched controls. The BAD patients were diagnosed as described earlier in Severinsen *et al.* (2006).

Selection of genetic markers

Genomic DNA was extracted from whole blood from patients and controls using standard methods. The selection criteria for the 13 microsatellite loci were based on previous positive findings (Degn *et al.*, 2001; Kalsi *et al.*, 2006).

Five SNPs were selected based on the positive findings in Kalsi *et al.* (2006) and four additional SNPs were chosen based on a functional approach. Of the four additional SNPs, two are located in the proximal promoter region (m6 and m11) and two are located within exons (m7 and m13) of the various Slynar transcripts. Most Slynar transcripts seem to be noncoding, however, we used the translate tool from ExPASy (*http://www.expasy.ch/tools/*) to predict protein sequences.

Marker m7 is located in the promoter region of Slynar_f and Slynar_a irrespective of reading frames. Marker m13 is located within exons of the Slynar_a, Slynar_c and Slynar_d transcripts. Protein prediction of the three transcripts show the possibility of m13 being a missense SNP in all three transcripts.

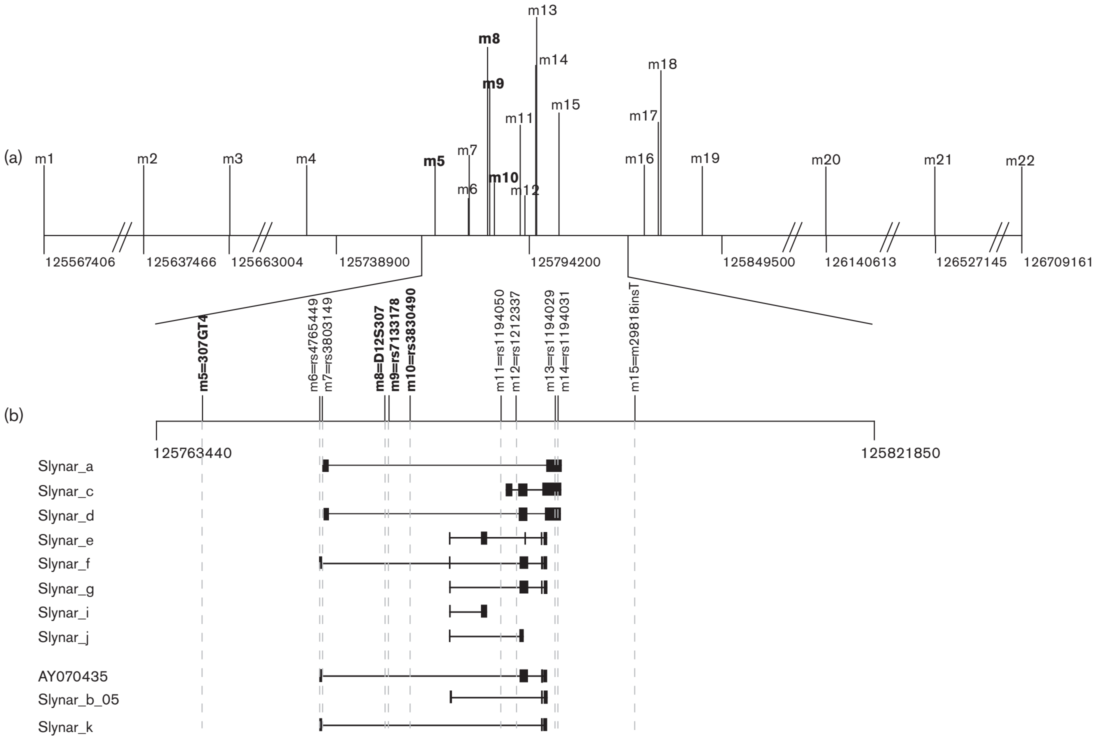
The SNP earlier denoted pufu in/del is referred to as rs3830490 or m10 in this article (see Fig. 1 and Table 1 for positions of genetic markers).

Analysis of genetic markers

Standard and multiplex PCR conditions were applied using a template of 36–40 ng genomic DNA in a total volume of 6–8 μl. Following PCR, nucleotides and primers were degraded using exonuclease 1 and shrimp alkaline phosphatase (GE Healthcare, Piscataway, New Jersey, USA) and the SNPs were genotyped according to the SNaPshot protocol (Applied Biosystems, Foster City, California, USA). SNPs and microsatellite PCRs were analyzed on an ABI 3100 Prism Genetic Analyzer (Applied Biosystems) and the genotyping data were analyzed using Genemapper software version 3.7 (Applied Biosystems). All data were independently checked by two experienced investigators. Any inconsistency led to regenotyping of the sample for that particular marker. The observed maximum proportion of missing data was 1% for microsatellites and 2% for SNPs. Owing to clustering of genotyping failures in one BAD case and two control individuals these were excluded in the subsequent analyses.

The Scottish samples were genotyped for m9 using the Sequenom platform (Sequenom MassARRAY System, Sequenom, Inc., San Diego, USA). PCR was carried out in a 384-well microtiter plate using 10 ng genomic DNA as template in a total volume of 5 μl. Following PCR, the amplicons were treated with shrimp alkaline phosphatase (Sequenom, Inc.). The extension reaction was carried out using the iPLEX Gold reaction mix (Sequenom, Inc.) and thereafter desalted by adding resin. The extension products were spotted onto a 384 SpectroCHIP Array using a Nanodispenser, and analyzed by the MassARRAY analyzer compact. We performed data analysis using the MassARRAY Typer software (version 4.0). The call rate

Fig. 1



(a) The position of the 22 selected markers (denoted as m1–m22) in a 1.14 mb region on chromosome 12q24 is shown according to their alignment with the March 2006 human reference sequence, which corresponds to the NCBI Build 36.1. Markers associated with the disease are printed in bold-faced type. (b) Enlargement of a 58-kb region harbouring the markers associated with the disease. The locations of selected predicted transcripts within this chromosomal region are shown. Predicted Sylnar transcripts from the April 2007 release of the AceView genemodels, the human AY070435 mRNAs from the Genbank, Slynar_b_05 and the novel identified Slynar transcript (Slynar_k) are shown. Predicted exons are shown as black boxes.

was 0.93 and no deviation from Hardy–Weinberg equilibrium (HWE) was observed ($P = 0.34$). Eighty-two samples were genotyped twice with a concordance rate of 100%. Twenty-seven samples (14 patients with BAD and 13 control individuals) were excluded because of failed genotyping. All information regarding PCR conditions and primer sequences are available on request.

Statistical methods

Single marker genotype-wise and allele-wise Fisher's exact association tests were performed. Permutation-based P values using 1e6 simulations were used for microsatellites as these are highly variable. Logistic regression was applied on significantly associated markers to assess their disease risk in the best fitting genetic model. The models considered, apart from the saturated and the null, were the dominant, additive and recessive models. Odds ratios (OR) with 95% confidence intervals (CI) were estimated. Haplotype analyses were performed only for SNPs by a sliding window approach with two, three and four consecutive markers using the score method of Schaid *et al.* (2002). The global score test statistics were evaluated by means of 1e6 simulations. Haplotype-specific scores were examined whenever the global test statistic was significant. Linkage disequilibrium (LD) in terms of r^2 was estimated for all pairs of SNP markers. LD between microsatellite marker m8 and SNP marker m9 were calculated using software for analysis and visualization of interallelic LD between multiallelic markers (Multiallelic Interallelic Disequilibrium Analysis Software) (Gaunt *et al.*, 2006). Using Multiallelic Interallelic Disequilibrium Analysis Software, LD is calculated for each allelic combination between all pairwise combinations of any type of loci.

Fisher's exact test for Hardy–Weinberg equilibrium HWE was carried out by performing 1e4 permutations in the

Table 1 **Genomic markers analyzed**

Marker	Marker name	Position[a]	Marker type	Genotyped in Kalsi *et al.* (2006)
1634GT2	m1	125567406	Microsatellite	+
AFMb337ZD5	m2	125637466	Microsatellite	+
1634tet	m3	125663004	Microsatellite	+
D12S1634	m4	125730415	Microsatellite	+
307GT4	m5	125767158	Microsatellite	+
rs4765449	m6	125776761	SNP	
rs3803149	m7	125776974	SNP	
D12S307	m8	125782285	Microsatellite	+
rs7133178	m9	125782861	SNP	+
rs3830490	m10	125784247	SNP	+
rs1194050	m11	125791734	SNP	
rs1212337	m12	125792925	SNP	+
rs1194029	m13	125796148	SNP	
rs1194031	m14	125796385	SNP	+
m29818insT	m15	125802773	SNP	+
X307CA1	m16	125827215	Microsatellite	+
X307CA2	m17	125831253	Microsatellite	+
D12SDK2	m18	125832036	Microsatellite	+
D12SDK1	m19	125843947	Microsatellite	+
D12S1658	m20	126140613	Microsatellite	+
D12S2075	m21	126527145	Microsatellite	+
D12S1675	m22	126709161	Microsatellite	

SNP, single-nucleotide polymorphism.

[a] The positions are according to the University of California Santa Cruz genome browser, March 2006 freeze (*http://genome.ucsc.edu/*).

GDA software (Lewis and Zaykin, 2001). Power calculations were performed using statistics from Long *et al.* (1997). The meta-analysis of m9 across the Danish, Scottish and UK samples consisted of a stratified logistic regression analysis carried out using Stata10 (StataCorp, College Station, Texas, USA). Genotypes for the m9 marker (rs7133178) in the UK samples were generated on basis of the allele frequencies reported in Kalsi *et al.* (2006) assuming HWE. Correction for multiple testing was considered by Hommel's method of controlling the family-wise error rate (Hommel, 1988), which is more powerful (Shaffer, 1995) than Bonferroni correction. All other analyses were performed using the *genetics* and *haplo.stats* packages in R (R Development Core Team, 2004). A significance level of 5% was chosen.

Analysis of Slynar transcripts

According to the AceView gene models, which is one of the gene tracks from the University of California Santa Cruz (UCSC) Genome Browser (*http://genome.ucsc.edu/*), 10 predicted Slynar transcripts exist in the April 2007 release of the program compared with five transcripts in the August 05 version. However, two of the transcripts from the April 2007 release are unspliced forms and will not be considered in this article. A comparison of the April 2007 and August 05 release of the program shows a high degree of similarity between parts of the predicted transcripts, although the genomic positions have changed.

The AceView program (*http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/*) developed at NCBI provides a strictly cDNA supported analysis of the human transcriptome and genes. For each new release, the models are improved by incorporating the latest cDNA data, and AceView seems to provide one of the most comprehensive and accurate representations of the entire human transcriptome (Thierry-Mieg and Thierry-Mieg, 2006).

To confirm the predicted Slynar transcripts and to identify possible novel transcripts, primers were designed to align within each of the known Slynar exons. There are several overlaps between exons in the different transcripts and combinations of these were tested. We used human brain, whole Marathon–Ready cDNA (Clontech, Mountain View, California, USA) as template and the PicoMaxx polymerase (Stratagene, La Jolla, California, USA) for amplification of the cDNA. If no PCR products were visible on the agarose gel, an additional PCR of 35 cycles was performed using the same primers. Then, if still no PCR products were visible, we concluded that the specific transcript was either nonexistent or only present in the cDNA at very low levels.

Using primers for Slynar_k and Slynar_b_05, human expression patterns were examined in a panel of selected tissues by quantitative real-time PCR (qRT-PCR) on cDNAs synthesized from poly A + mRNA or total RNA isolated from various human brain and peripheral tissues (Clontech). The forward and reverse primers for the Slynar_k transcript were 5′-CCGCAAATGTGACCCGC AATT-3′ and 5′-CTCTCCTCTGGCACGGAAAC-3′, respectively while the forward and reverse primers for the Slynar_b_05 transcript were 5′-CCAGATACGGGTACTG TTGTAACTC-3′ and 5′-GAAAACCACCAATGCAATC C-3′. First strand synthesis of cDNA was performed using the Taqman reverse transcription reagents (Applied Biosystems), according to the manufacturer's recommendation. After synthesis of cDNA, qRT-PCR was performed using the iTaq SYBR Green Supermix w/ROX (BioRad, Hercules, California, USA) according to the manufacturer's recommendation. Quantitative PCR measurements were collected on the DNA Engine Opticon (MJ Research, Waltham, Massachusetts, USA), and subsequently analyzed by applying the $2^{-\Delta\Delta C}$T method (Winer *et al.*, 1999; Schmittgen *et al.*, 2000; Livak and Schmittgen, 2001). In short, the relative expression level of each cDNA was calculated by normalizing to the expression levels of peptidylprolyl isomerase A in the sample, and set relative to the mean normalized expression level of the testis sample. Glyceraldehyde-3-phosphate dehydrogenase and 18s rRNA were used as negative controls (data not shown).

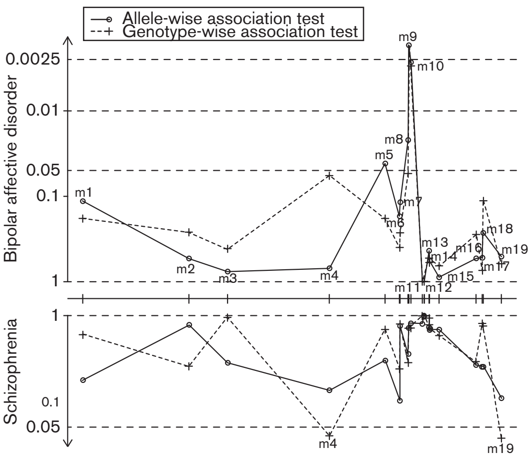
Results

Single marker analysis, bipolar affective disorder

Two markers, m9 and m10 were significantly associated with BAD for both genotype-based ($P = 0.002$ and 0.003, respectively) and allele-based analyses ($P = 0.002$ and 0.003, respectively). The minor allele (T in m9 and G in m10) was overrepresented among cases for both markers: 14% of the BAD patients carried the T-allele at m9 compared with 7% of the controls and 14% of the BAD patients carried the G-allele at m10 compared with 8% of the controls. Two other markers (m5 and m8) showed allelic association with disease status ($P = 0.04$ and 0.02, respectively).

Figure 2 shows the P values on a base-10 logarithmic scale for the genotype-wise and allele-wise association tests plotted against chromosomal positions. The genotype and allele counts for the genotyped SNPs are shown in (Table 2 in the Supplementary material). The impact of m9 and m10 on disease risk was assessed by logistic regression. The fully saturated genotype model (genotype-based association) were superior to the null model but did not fit significantly better than the corresponding additive and dominant models. A recessive genetic model was not supported. Based on Akaike's Information Criterion the additive genetic model was chosen and the OR for carriers of an additional minor allele was found to be 2.0 (95% CI: 1.3–3.2) for m9 and 2.0 (95% CI: 1.3–3.1) for m10. The additive effect is multiplicative on the OR scale (exponentiated difference of log odds). Thus, the OR for homozygous carriers of the minor allele is the square of these ORs', that is, OR = 4.

Fig. 2



P values for the genotype-wise and allele-wise association tests in the Danish bipolar and schizophrenia sample plotted on a base-10 logarithmic scale against chromosomal positions.

Replication and refinement of association between bipolar disorder and the Slynar locus was the main purpose of the study. This corresponds to a main null hypothesis saying that none of the 11 markers within the Slynar region (m5-m15) are associated with BAD. The allelic associations with m9 and m10 both survive correction for this family of tests by Hommel's procedure (Hommel, 1988) and the corrected P values were 0.017 and 0.027, respectively. If the family of tests is broadened to be all 22 allelic tests then the corrected P values are 0.036 and 0.057, and only m9 is still significantly associated.

Furthermore, m9 showed association with BAD in the Scottish sample ($P_{\text{genotype}} = 0.03$) and in the combined Danish and Scottish sample ($P_{\text{allele}} = 0.004$ and $P_{\text{genotype}} = 0.008$). Similar to the Danish and the UK sample (Kalsi *et al.*, 2006), the minor allele was overrepresented in cases compared with controls in the Scottish sample.

In the combined Danish, Scottish and UK sample (918 patients with BAD and 946 controls in total), a meta-analysis of m9 using the additive genetic model showed an OR of 1.5 (95% CI: 1.2–1.9) for carriers of the minor allele ($P = 0.0003$). Correspondingly, the OR for homozygous carriers of the minor allele was OR = 2.2.

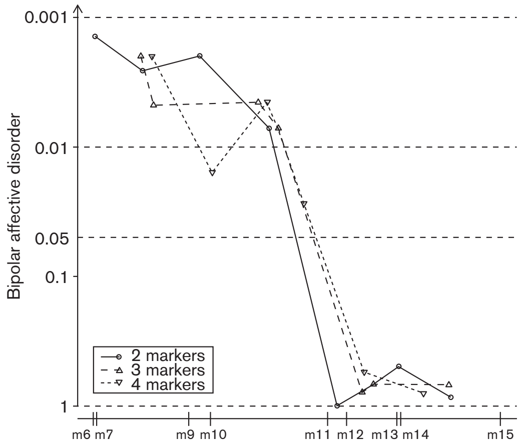
Single marker analysis, schizophrenia

Three microsatellites, m4, m19 and m22 were significantly associated with SZ in a genotype-based test for association ($P = 0.04$, 0.04 and 0.03, respectively). However, these markers were not associated with SZ in allele-based tests (Fig. 2, results not shown for m22).

Haplotype analysis

The distribution of several two and three and four-marker haplotypes were significantly different between BAD cases and controls with P values below 0.01 (Fig. 3). The most significantly associated two-marker haplotype included m6-m7 ($P = 0.001$). This haplotype association was primarily caused by differences in the frequencies of two

Fig. 3



P values from two and three and four-marker haplotype association analysis on single-nucleotide polymorphisms only in the Danish bipolar sample plotted on a base-10 logarithmic scale versus chromosomal positions.

haplotypes (C-C and A-G). The C-C haplotype seems to be a risk haplotype ($P_{\text{local}} = 0.0024$) with frequencies of 0.139 and 0.076 among BAD cases and controls, respectively, whereas the A-G haplotype seems to be a protective haplotype ($P_{\text{local}} = 0.0054$) with frequencies of 0 and 0.0211 among BAD cases and controls, respectively. The most significantly associated three and four-marker haplotypes also included m6-m7. However, several of the remaining significantly associated haplotypes involved marker m9 and m10 supporting the results from the single marker analysis.

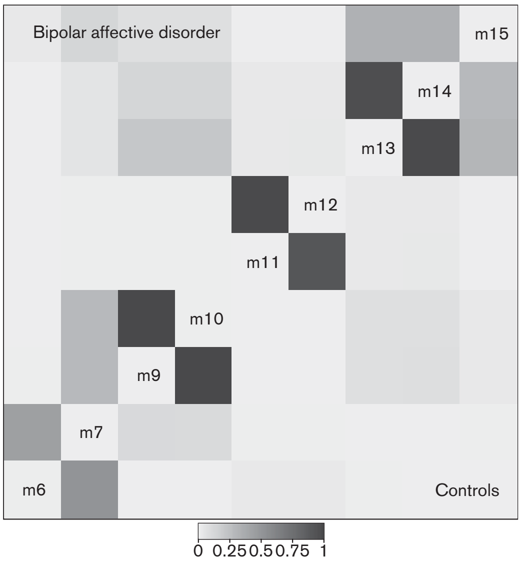
A single three-marker haplotype consisting of m6-m7-m9 and a four-marker haplotype consisting of m6-m7-m9-m10 seemed to be associated with SZ ($P = 0.03$ and $P = 0.03$, respectively). No two-marker haplotypes were associated with SZ (results not shown).

Linkage disequilibrium

The LD pattern for pairs of SNP markers, measured in terms of r^2 is shown in Fig. 4 for BAD cases and controls. It should be noted that m9 and m10 are in strong LD ($r^2 = 0.99$ in cases and $r^2 = 0.98$ in controls). Haplotypes of microsatellite marker m8 and SNP marker m9 showing the highest difference in frequencies between cases and controls (13 and 7%, respectively) were also in strong LD with one another ($r^2 = 0.95$) (results not shown).

As expected, LD was strongest among markers with dense distance. LD among BAD cases was slightly stronger than among control individuals and SZ cases (LD results for SZ are not shown).

Fig. 4



Pairwise linkage disequilibrium for single-nucleotide polymorphisms in the Danish bipolar and control sample in terms of r^2 shown in increasing greyscale colours.

Hardy–Weinberg equilibrium

A few markers deviated from HWE, all of which were microsatellites. In the BAD sample m18 had a *P* value of 0.04. In the SZ sample m19 and m22 had *P* values below 5% ($P = 0.05$ and 0.02, respectively). The largest deviation was observed for m5 in the control group with a *P* value of 0.002.

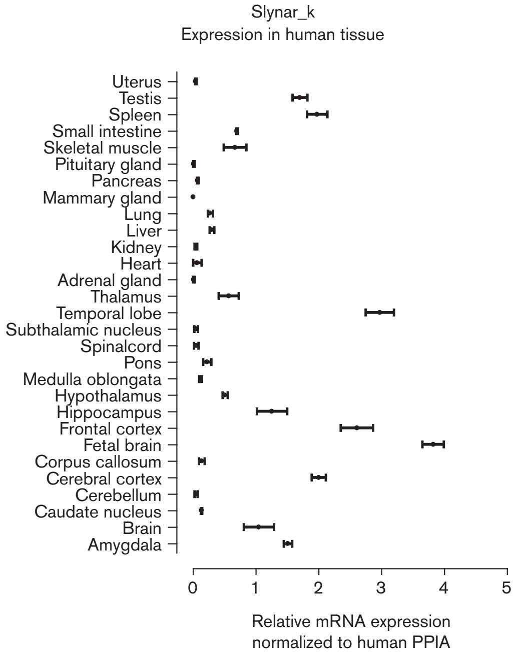
Analysis of Slynar transcripts

We investigated the various Slynar transcripts and confirmed the existence of the splicing variant Slynar_b from the old August 2005 version of Aceview, denoted Slynar_b_05 in this article. The mRNA sequence of Slynar_b_05 comprises 512 bp. We found the mRNA sequence of the transcript to be longer than predicted by the August 2005 version of AceView. None of the April 2007 transcripts were 100% identical to Slynar_b_05, however. Nevertheless, with exception of 41 bp in the 5′end of Slynar_b_05, a sequence alignment between Slynar_b_05 and Slynar_f reveals 100% identity between Slynar_b_05 and exons 3 and 4 from Slynar_f. We furthermore identified a novel transcript in the human brain which we denoted Slynar_k. A sequence alignment of the novel identified Slynar_k transcript showed close similarity to exons 1, 3 and 4 from Slynar_f. Both identified transcripts, Slynar_b_05 and Slynar_k, were sequenced to confirm their identity. However, despite several attempts using different combinations of primers, we were unable to confirm the existence of any other Slynar AceView transcripts.

Figure 1 shows the predicted Slynar transcripts from the April 2007 release of the AceView gene models including the novel identified Slynar transcripts, Slynar_k and the Slynar_b_05. However, the position of Slynar_b_05 is based on sequence alignment with transcripts from the April 2007 version of AceView, as the genomic sequences positions have changed.

Both Slynar_b_05 and Slynar_k were further analyzed in a wide range of human tissues using qRT-PCR. Human Slynar_b_05 seemed to be most abundant in testis (results not shown). No expression of Slynar_b_05 was detected in other human tissues by the qRT-PCR, however, we cannot exclude that it is expressed at very low levels in some tissues, here including brain. In contrast, the relative expression levels of Slynar_k were found highest in tissues from foetal brain, frontal cortex, cerebral cortex, temporal lobe, testis and spleen (Fig. 5). In fact, Slynar_k was generally expressed in more tissues including the testis, however, the expression in tissues from the central nervous system was more pronounced than the expression in peripheral tissues with lowest levels in kidney, pancreas and heart.

Fig. 5



Expression of Slynar_k in various human brain and peripheral tissues. Triplicate quantitative real-time-PCR data is presented as mean ± SEM. PPIA, peptidylprolyl isomerase A.

Discussion

The results presented in this study further support the presence of a susceptibility locus for BAD on chromosome 12q24.3. Of the 1.14 Mb surveyed, allele-wise and genotype-wise test for association implicate a 50 kb region within the Slynar locus. The two most significantly associated markers, m9 and m10 from this study, were also associated with BAD in the UK cohort and have never previously been genotyped in the Danish sample (Kalsi *et al.*, 2006). The two associated markers were highly correlated ($r^2 > 0.98$). Both markers survive Hommel's correction for multiple testing (Hommel, 1988).

The power to detect an OR = 2 for carriers of one minor allele in an additive model was found to be 77% under assumptions of a disease prevalence of 1%, and a 14% frequency of the risk allele, as observed for the most significantly associated markers in the Danish BAD sample.

To replicate the association of the most significantly associated marker in the Danish sample, m9 was also analyzed in a Scottish sample. The association was confirmed, thereby supporting the Slynar locus as a susceptibility locus for BAD.

As m9 was associated with BAD in the three samples, we also performed a meta-analysis showing an OR = 1.5 for carriers of one minor allele and an OR = 2.2 for carriers of two minor alleles.

No markers in the genome-wide association (GWA) study of the Wellcome Trust Case Control Consortium (2007) or in the study by Sklar *et al.* (2008) show any significant association with BAD within 12q24. This also applies to the SNP marker, rs1706509 included on the Affymetric GeneChip Human Mapping 500K Array Set, which is located in relative proximity of the BAD associated marker, m9 (rs7133178) of this study. Despite the relative short distance between rs7133178 and rs1706509 (3208 base pairs), they are in very low LD, $r^2 = 0.005$, which could explain why there is no significant association of BAD with rs1706509 in the two GWA studies.

Three SNPs from our study (m6, m9 and m11) are however included on the Illumina HumanHap550 chip. Unfortunately, the only bipolar GWA study using the Illumina chip also used DNA pooling, a method known to reduce the power to detect genetic association (Baum *et al.*, 2008).

According to Haploview, the Slynar locus (m5-m15) is defined by three haplotype blocks, and 14 tagSNPs capture 34 SNPs to obtain full coverage ($r^2 \geq 0.8$). Our study comprised six tagSNPs (m6, m7, m9, m11, m12, m13) within this chromosome region. These markers represented each of the defined haplotype blocks and captured 21 SNPs at $r^2 \geq 0.8$. In addition, we included two microsatellites (m5, m8) and three additional SNPs (m10, m14, m15) within the Slynar locus, indicating that most of the common genetic variation has been assessed by the combined set of markers.

As SNPs located in the promoter region of a gene may influence the gene expression by changing the ability of transcription factors to bind to the DNA sequence, we analyzed the impact of a potential promoter SNP (m9) on potential binding sites for transcription factors using the program MatInspector (*www.genomatix.de*) (Quandt *et al.*, 1995; Werner, 2000). The major allele in m9 may influence transcription by leading to an alternative binding site for a transcription factor (BRN2/POUF3) (showing a high core and matrix similarity). This transcription factor belongs to a large family of transcription factors predominantly expressed in the central nervous system with a possible role in neurogenesis (Schreiber *et al.*, 1993; Atanasoski *et al.*, 1995).

Furthermore, this study replicates the significant allele-based association of both m5 and m8 by Kalsi *et al.* (2006). The most significantly associated marker in Kalsi *et al.* (2006) m19, was not associated with BAD in this study. It is however notable that this marker was associated with SZ in the Danish sample.

In general, the results from the haplotype analysis in the BAD sample supported the results from the allele-wise and genotype-wise test for association. But, independently of the single marker analyses the most significantly associated two, three, and four-marker haplotypes included markers m6 and m7. The associated three and four-marker haplotypes in the SZ sample also included m6 and m7.

In SZ none of the nominally significant associations observed in the single marker and haplotype analyses could withstand correction for multiple testing. Thus, the results do not suggest that the Slynar locus is a common susceptibility locus for BAD and SZ. However, further studies are needed to reject this hypothesis.

We furthermore investigated the possible effect of intragenic SNPs (m6, m7, m11, m12, m13 and m14) on splicing using the programs ESE-finder release 3.0 (*http://rulai.cshl.edu/tools/ESE*) (Cartegni *et al.*, 2003; Smith *et al.*, 2006), FAS-ESS web server (*http://genes.mit.edu/fas-ess/*) (Wang *et al.*, 2004) and RESCUE-ESE web server (*http://genes.mit.edu/burgelab/rescue-ese/*) (Fairbrother *et al.*, 2002; Yeo *et al.*, 2004). However, the analyzed SNPs did not show any potential effect on splicing. We furthermore analyzed the possible effect of 3′ untranslated region SNPs (m13 and m14) on microRNA binding using the miRBase Target Release version 1 (*http://microrna.sanger.ac.uk/*) and did not find any differential effects of the alleles.

A promising candidate gene in the genomic region showing association with BAD is denoted Slynar in the AceView database. Most of the Slynar transcripts seem to be noncoding. The best predicted proteins would be around 80 amino acids long except from Slynar_a which has a predicted protein of 216 amino acids and a good protein coding score according to the AceView database. We were not able to identify the Slynar_a transcript, however, but we confirmed the existence of the Slynar_b_05 transcript and identified a novel Slynar transcript, Slynar_k, in human brain cDNA. The longest predicted proteins within Slynar_b_05 and Slynar_k are 99 and 111 amino acids, respectively. The length of the predicted Slynar_k protein was obtained using the translate tool from ExPASy (*http://www.expasy.ch/tools/*). We have not tested whether these possible proteins are in fact expressed. Using gene-specific primers for Slynar_k and Slynar_b_05, human expression patterns were examined using qRT-PCR on cDNAs from various brain and peripheral tissues, and showed the relative expression level of Slynar_k to be high in several different brain tissues.

The chromosome region confined by our most distant markers m1 (1634GT2) and m22 (D12S1675) is very gene poor and only harbours six predicted genes (UCSC gene predictions). More locally within the region surrounding our most significantly associated single markers (m9 and m10) there are only four predicted UCSC genes. Two of these predicted genes, BC039096 and CR615184, show high similarity to several exons in the Slynar transcripts. As CR615184 and BC039096 have been predicted as UCSC genes relatively recently, we have not included them in our analyses of transcripts. The other two predicted genes within the region surrounding m9 are antisense to Slynar and might have a role in regulating the expression of Slynar.

Noncoding RNA has been shown to regulate almost every level of gene expression (Amaral and Mattick, 2008). We performed a BLAST search to identify small antisense sequences (20–25 bp) with perfect matches to the various Slynar transcripts. The results pointed towards two genes (Ral guanine nucleotide dissociation stimulator-like 1 and Ring finger protein 141) hypothetically able to be regulated by most of the Slynar transcripts.

The function of Slynar still seems unknown, however, and further functional studies are needed to clarify the function of the gene and the importance of this gene in BAD.

In conclusion, the exact replication of markers associated with disease status supports 12q24.3 as a region of functional importance in the pathogenesis of BAD. One marker in the Slynar locus seemed to be associated with BAD in three independent samples. As no SNPs analyzed in the recent (nonpooling) GWA studies of BAD are good proxies for this marker, the present results confirm the importance of focused genotyping.

Acknowledgements

The authors acknowledge skillful assistance of A. Hedemand. The study was supported by the JASCHA foundation, the 'Pulje til Styrkelse af Psykiatrisk Forskning i Aarhus Amt', the Augustinus Foundation and by the Stanley Medical Research Institute. D.B. and W.M. were supported by grants from the Chief Scientist Office of the Scottish Government and The Wellcome Trust, UK.

Conflicts of interest: none declared.

References

Amaral PP, Mattick JS (2008). Noncoding RNA in development. *Mamm Genome* **19**:454–492.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders*. Washington: American Psychiatric Association.

Atanasoski S, Toldo SS, Malipiero U, Schreiber E, Fries R, Fontana A (1995). Isolation of the human genomic brain-2/N-Oct 3 gene (POUF3) and assignment to chromosome 6q16. *Genomics* **26**:272–280.

Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, *et al.* (2008). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* **13**:197–207.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**:3568–3571.

Curtis D, Kalsi G, Brynjolfsson J, McInnis M, O'Neill J, Smyth C, *et al.* (2003). Genome scan of pedigrees multiply affected with bipolar disorder provides further support for the presence of a susceptibility locus on chromosome 12q23–q24, and suggests the presence of additional loci on 1p and 1q. *Psychiatr Genet* **13**:77–84.

Degn B, Lundorf MD, Wang A, Vang M, Mors O, Kruse TA, Ewald H (2001). Further evidence for a bipolar risk gene on chromosome 12q24 suggested by investigation of haplotype sharing and allelic association in patients from the Faroe Islands. *Mol Psychiatry* **6**:450–455.

Ewald H, Degn B, Mors O, Kruse TA (1998). Significant linkage between bipolar affective disorder and chromosome 12q24. *Psychiatr Genet* **8**:131–140.

Ewald H, Flint T, Kruse TA, Mors O (2002). A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24.3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13.3. *Mol Psychiatry* **7**:734–744.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* **297**:1007–1013.

Gaunt TR, Rodriguez S, Zapata C, Day IN (2006). MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. *BMC Bioinformatics* **7**:227.

Hommel G (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**:383–386.

Jones I, Jacobsen N, Green EK, Elvidge GP, Owen MJ, Craddock N (2002). Evidence for familial cosegregation of major affective disorder and genetic markers flanking the gene for Darier's disease. *Mol Psychiatry* **7**:424–427.

Kalsi G, McQuillin A, Degn B, Lundorf MD, Bass NJ, Lawrence J, *et al.* (2006). Identification of the Slynar gene (AY070435) and related brain expressed sequences as a candidate gene for susceptibility to affective disorders through allelic and haplotypic association with bipolar disorder on chromosome 12q24. *Am J Psychiatry* **163**:1767–1776.

Kato T (2007). Molecular genetics of bipolar disorder and depression. *Psychiatry Clin Neurosci* **61**:3–19.

Lewis PO, Zaykin D (2001). Genetic data analysis: computer program for the analysis of allelic data. Version 1.0 (d16c).

Livak KJ, Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**:402–408.

Long AD, Grote MN, Langley CH (1997). Genetic analysis of complex diseases. *Science* **275**:1328.

Lyons-Warren A, Chang JJ, Balkissoon R, Kamiya A, Garant M, Nurnberger J, *et al.* (2005). Evidence of association between bipolar disorder and Citron on chromosome 12q24. *Mol Psychiatry* **10**:807–809.

Morissette J, Villeneuve A, Bordeleau L, Rochette D, Laberge C, Gagne B, *et al.* (1999). Genome-wide search for linkage of bipolar affective disorders in a very large pedigree derived from a homogeneous population in quebec points to a locus of major effect on chromosome 12q23-q24. *Am J Med Genet* **88**:567–587.

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res* **23**:4878–4884.

R Development Core Team (2004). R: a language and environment for Statistical Computing, Vienna, Austria. R Foundation for statistical computing.

Reif A, Herterich S, Strobel A, Ehlis AC, Saur D, Jacob CP, *et al.* (2006). A neuronal nitric oxide synthase (NOS-I) haplotype associated with schizophrenia modifies prefrontal cortex function. *Mol Psychiatry* **11**:286–300.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**:425–434.

Schmittgen TD, Zakrajsek BA, Mills AG, Gorn V, Singer MJ, Reed MW (2000). Quantitative reverse transcription-polymerase chain reaction to study mRNA decay: comparison of endpoint and real-time methods. *Anal Biochem* **285**:194–204.

Schreiber E, Tobler A, Malipiero U, Schaffner W, Fontana A (1993). cDNA cloning of human N-Oct3, a nervous-system specific POU domain transcription factor binding to the octamer DNA motif. *Nucleic Acids Res* **21**:253–258.

Severinsen JE, Bjarkam CR, Kiaer-Larsen S, Olsen IM, Nielsen MM, Blechingberg J, *et al.* (2006). Evidence implicating BRD1 with brain development and susceptibility to both schizophrenia and bipolar affective disorder. *Mol Psychiatry* **11**:1126–1138.

Shaffer JP (1995). Multiple hypothesis testing. *Annu Rev Psychol* **46**:561–584.

Shink E, Harvey M, Tremblay M, Gagne B, Belleau P, Raymond C, *et al.* (2005a). Analysis of microsatellite markers and single nucleotide polymorphisms in candidate genes for susceptibility to bipolar affective disorder in the chromosome 12Q24.31 region. *Am J Med Genet B Neuropsychiatr Genet* **135B**:50–58.

Shink E, Morissette J, Sherrington R, Barden N (2005b). A genome-wide scan points to a susceptibility locus for bipolar disorder on chromosome 12. *Mol Psychiatry* **10**:545–552.

Shinkai T, Ohmori O, Hori H, Nakamura J (2002). Allelic association of the neuronal nitric oxide synthase (NOS1) gene with schizophrenia. *Mol Psychiatry* **7**:560–563.

Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, *et al.* (2008). Whole-genome association study of bipolar disorder. *Mol Psychiatry* **13**:558–569.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* **15**:2490–2508.

Thierry-Mieg D, Thierry-Mieg J (2006). AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7** (**Suppl 1**):S12–S14.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* **119**:831–845.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447**:661–678.

Werner T (2000). Computer-assisted analysis of transcription control regions. Matinspector and other programs. *Methods Mol Biol* **132**:337–349.

Winer J, Jung CK, Shackel I, Williams PM (1999). Development and validation of real-time quantitative reverse transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes in vitro. *Anal Biochem* **270**:41–49.

World Health Organization (1993). The ICD10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research.

World Health Organization (1998). Diagnosis and clinical measurement in psychiatry. A reference manual for SCAN.

Yeo G, Hoon S, Venkatesh B, Burge CB (2004). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci U S A* **101**:15700–15705.

Zubenko GS, Hughes HB III, Stiffler JS (1999). Neurobiological correlates of a putative risk allele for Alzheimer's disease on chromosome 12q. *Neurology* **52**:725–732.

6.5.1 Paper 5 - supplementary table

Genetic markers			Genotype counts: MM/Mm/mm (freq.)			Allele counts: M/m (freq.)		
ID	Name	Position[a]	Bipolar	Schizophrenia	Controls	Bipolar	Schizophrenia	Controls
rs4765449	m6	125776761	120/42/3 (0.73/0.25/0.02)	149/49/4 (0.74/0.24/0.02)	206/92/10 (0.67/0.30/0.03)	282/48 (0.85/0.15)	347/57 (0.86/0.14)	504/112 (0.82/0.18)
rs3803149	m7	125776974	88/60/17 (0.53/0.36/0.10)	126/63/15 (0.62/0.31/0.07)	183/103/21 (0.60/0.34/0.07)	236/94 (0.72/0.28)	315/93 (0.77/0.23)	469/145 (0.76/0.24)
rs7133178	m9	125782861	122/38/4 (0.74/0.23/0.02)	173/29/2 (0.85/0.14/0.01)	264/44/1 (0.85/0.14/<0.01)	282/46 (0.86/0.14)	375/33 (0.92/0.08)	572/46 (0.93/0.07)
rs3830490	m10	125784247	123/38/4 (0.75/0.23/0.02	171/29/2 (0.85/0.14/0.01)	263/45/1 (0.85/0.15/<0.01)	284/46 (0.86/0.14)	371/33 (0.92/0.08)	571/47 (0.92/0.08)
rs1194050	m11	125791734	156/5/0 (0.97/0.03/0)	197/6/0 (0.97/0.03/0)	298/9/1 (0.97/0.03/<0.01)	317/5 (0.98/0.02)	400/6 (0.99/0.01)	605/11 (0.98/0.02)
rs1212337	m12	125792925	159/5/0 (0.97/0.03/0)	197/6/0 (0.97/0.03/0)	297/10/0 (0.97/0.03/0)	323/5 (0.98/0.02)	400/6 (0.99/0.01)	604/10 (0.98/0.02)
rs1194029	m13	125796148	61/84/20 (0.37/0.51/0.12)	87/93/22 (0.43/0.46/0.11)	130/142/37 (0.42/0.46/0.12)	206/124 (0.62/0.38)	267/137 (0.66/0.34)	402/216 (0.65/0.35)
rs1194031	m14	125796385	64/81/20 (0.39/0.49/0.12)	89/93/21 (0.44/0.46/0.10)	132/136/38 (0.43/0.44/0.12)	209/121 (0.63/0.37)	271/135 (0.67/0.33)	400/212 (0.65/0.35)
m29818insT	m15	125802773	65/77/23 (0.39/0.47/0.14)	75/97/31 (0.37/0.48/0.15)	113/157/37 (0.37/0.51/0.12)	207/123 (0.63/0.37)	247/159 (0.61/0.39)	383/231 (0.62/0.38)

a) The positions are according to the UCSC genome browser, March 2006 freeze (http://genome.ucsc.edu/).

6.6 Paper 6[35]

LandScape: A Simple Method to Aggregate P-Values and Other Stochastic Variables Without a Priori Grouping

Carsten Wiuf[1,*,†], Jonatan Schaumburg-Müller Pallesen[2,3,4,5,†], Leslie Foldager[3,4,5,6] and Jakob Grove[2,3,4,5]

[1] Department of Mathematical Science, University of Copenhagen, Copenhagen, Denmark
[2] Department of Biomedicine, Aarhus University, Aarhus, Denmark
[3] *i*PSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus and Copenhagen, Denmark
[4] *i*SEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark
[5] Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
[6] Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Risskov, Denmark
[*] Corresponding author: wiuf@math.ku.dk [†] Joint first authors

Aarhus University, 31 March 2014

Abstract

Motivation: In many areas of science it is custom to perform many, potentially millions, of tests simultaneously. To control the number of false discoveries different strategies are used, such as correction for multiple testing and grouping of tests based on a priori criteria. It is, however, not straightforward to choose grouping criteria. Methods that summarize, or aggregate, test statistics or p-values, without relying on a priori criteria, is therefore desirable.

Results: We present a simple method to aggregate sequentially ordered stochastic variables, such as test statistics or p-values, into fewer variables without assuming a priori defined groups. We provide different ways to evaluate the significance of the aggregated variables based on theoretical considerations, using ideas from random walk theory, and bootstrap techniques.

Availability and Implementation: Implementations of the method in R and Python are available from the authors on request.

Proofs: All proofs of statements in the main text are given in the Appendix.

[35] Manuscript in preparation.

1 Introduction

Today we frequently face the situation of performing millions of statistical tests simultaneously. To control the number of false discoveries, it is standard to adjust for multiple testing. However, this might reduce the power substantially, in particular, if the tests are strongly correlated. One solution is to calculate an "effective number of independent tests"; an idea due to Cheverud (2001) who used the eigenvalues of a trait correlation matrix to estimate the effective number of independent traits. The idea was later applied in the context of linkage disequilibrium (Nyholt, 2004). Another idea is to summarize test statistics across a priori defined groups. A combined test value is calculated from the observed variables in each group. In this way fewer tests are performed and dependencies might be removed or diminished.

Much of this work has centred on an observation by Fisher to aggregate p-values: If k independent tests are performed with p-values $p_1,\ldots,p_k$, then $-2\sum_i \log p_i \sim \chi^2(2k)$ (Fisher, 1932). Other similar methods include Stouffer et al. (1949) and Simes (1986). Importantly, Fisher's aggregated statistic can be statistically significant while none of the p-values individually are. Thus, it is possible to detect a combined effect that does not show in the individual tests.

The distribution of Fisher's statistic relies on the assumption of independence. If the p-values are dependent then the test may be anti-conservative. Another potential problem is that the groups (k tests is one group) are defined a priori. In association mapping, the test statistics might be aggregated over more or less arbitrary (sliding) windows or grouped by being in the same gene. However, gene size varies considerably. The markers we seek to identify might only be in a small part of a gene or across a gene boundary. Thus, the gene might not be the appropriate unit to work with.

We propose a method that summarize sequentially ordered test statistics without a priori grouping. Sequentially ordered tests often occur in genomics, association mapping, and time-series analysis. We develop a procedure, inspired by Random Walk theory (Karlin and Altschul, 1990; Karlin and Dembo, 1992), to combine a sequence of values into a single value without relying on a specific (a priori) grouping. The method crawls along the sequence searching for a stretch of consecutive values that *jointly* have a good "score". If the variables are independent then theory predicts the approximate distribution of the aggregated value. When this assumption is not fulfilled, we use bootstrap techniques to approximate the distribution.

All proofs are provided in the Appendix.

2 Aggregation of variables

2.1 A motivating example

Consider an ordered sequence of random variables Z_k, where k denotes the position. We think of Z_k as (a transformation of) a test statistic or a p-value and imagine a test is conducted for each position. Let $Z_k = +1$ if the k-th test is significant at level α or otherwise let $Z_k = -1$. A long interval of mainly $+1$ may be considered unlikely and indicative of deviance from the null hypothesis. We provide a definition of such segments and an algorithm to find them. Specifically, we identify intervals $[n,m]$ (called maximal segments), such that the partial sums $U_{nk} = \sum_{i=n}^{k} Z_i$ and $U_{km} = \sum_{i=k}^{m} Z_i$ are positive for all $n \leq k \leq m$, and such that $[n,m]$ is as large as possible, see Fig. 1 for an illustration.

In Section 2.2 we describe mathematically how to construct the maximal segments. They fall in two classes, *dependent* and *independent* segments, which we characterize in Section 2.3. The score of a dependent segment is constrained by the score of the previous independent segment; in particular, it must be smaller.

2.2 Segments and scores

Let $\mathbb{K}$ be a finite or infinite set of consecutive positive integers starting at 1. We call such a set an *index set*. The length of $\mathbb{K}$, that is, the number of elements in $\mathbb{K}$, is denoted $|\mathbb{K}|$ and it may be finite or infinite. Let Z_k, $k \in \mathbb{K}$, be a sequence of random variables and let U_{nm} be the partial sums:

$$U_{nm} = \sum_{k=n}^{m} Z_k, \quad n \leq m \in \mathbb{K}. \tag{2.1}$$

If $m < n$, we take the partial sum to be zero.

Definition 2.2. *A* segment *is a closed interval $[n,m]$ such that $U_{nk} > 0$ and $U_{km} > 0$ for all $k \in [n,m]$, where we allow m to be infinite. A* maximal segment *is a segment $[n,m]$ such that there is not another segment containing it. The* score *of a maximal segment $[n,m]$ is the partial sum U_{nm}.*

A (maximal) segment is always non-empty. Two different maximal segments, $[n_1,m_1]$ and $[n_2,m_2]$ are always disjoint. Assume conversely that $n_1 \leq n_2 \leq m_1 \leq m_2$. Then for $n_1 \leq k < n_2$, we have $U_{n_1,k} > 0$, as $[n_1,m_1]$ is a segment, and $U_{k,m_2} = U_{k,m_1} + U_{m_1+1,m_2} > 0$, since both $[n_1,m_1]$ and $[n_2,m_2]$ are segments, and similarly for $n_2 \leq k \leq m_1$ and $m_1 < k \leq m_2$. Thus $U_{n_1,k} > 0$ and $U_{k,m_2} > 0$ for all $k \in [n_1,m_2]$, which contradict the maximality of $[n_1,m_1]$ and $[n_2,m_2]$.

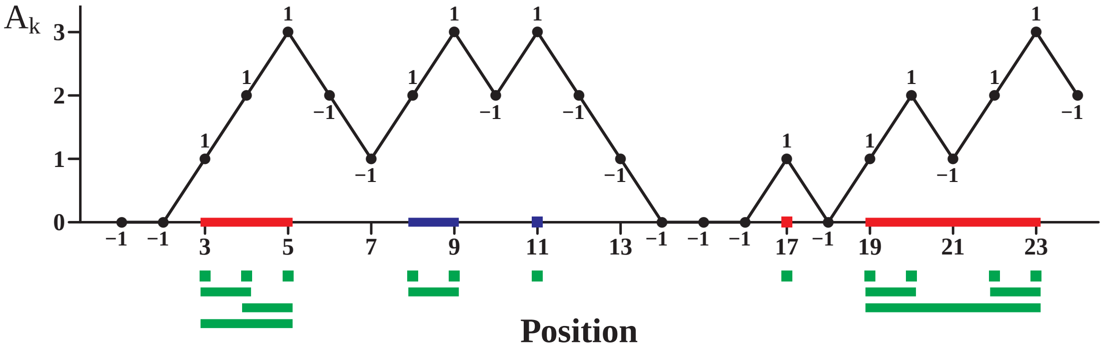
It follows that each position, $k \in \mathbb{K}$, is in at most one maximal segment. Hence, we have the following:

Theorem 2.3. *Let $Z_k, k \in \mathbb{K}$, be a sequence of random variables. Then there is a unique sequence of disjoint maximal segments $J_i = [n_i,m_i]$, $i \in \mathbb{M}$, containing all maximal segments. That is, if I is a maximal segment, then $I = J_i$ for some $i \in \mathbb{M}$.*

The sequence of segments $J_i, i \in \mathbb{M}$, is said to be *maximal*. If $[n,m]$ is a maximal segment then according to Definition 2.2, $Z_n, Z_m > 0$ and $Z_{n-1}, Z_{m+1} \leq 0$. If $Z_k \geq 0$ for all $k \in \mathbb{K}$, then there is at most one segment which also is maximal. If $Z_k > 0$ for at least one k then there is at least one segment, otherwise there are none. Thus, a natural requirement is that the random variables Z_k can take positive as well as negative values.

In the following, if $J_i = [n_i,m_i]$, $i \in \mathbb{M}$, is a sequence of maximal segments, we assume that n_i is increasing in i. This can always be achieved, potentially by reordering the segments.

Figure 1



Shown is a sequence of values $Z_k \in \{-1,1\}$ with positive values given above and negative below the points. The "landscape" is the accumulated sequence $A_k = \max\{0, Z_k + A_{k-1}\}$, see (2.4). The green bars show all segments, that is, intervals $[n,m]$ such that $U_{nk} > 0, U_{km} > 0$ (Definition 2.2), for example, $[3,4]$ is a segment. Maximal segments are indicated by coloured bars on the x-axis: red are independent segments and blue are dependent segments (Definition 2.8). A dependent segment starts when A_k increases after decreasing to a non-zero value.

2.3 Independent and dependent segments

In this section we present an algorithm to find all maximal segments. In the process the concept of independent and dependent segments will be introduced. Implementations of the algorithm in Python and R are available from the authors on request.

Formally, we let $A_0 = 0$ and define the accumulated sums by

$$A_k = \max\{0, Z_k + A_{k-1}\}, \quad k \in \mathbb{K}. \tag{2.4}$$

If $\mathbb{K}$ is finite, we put $A_{|\mathbb{K}|+1} = 0$. Define the start points (s_{i0}) and termination points (t_{i0}) by

$$\begin{aligned} s_{i0} &= \min\{k \in \mathbb{K} \mid k > t_{i-1,0}, A_k > 0\}, \\ t_{i0} &= \min\{k \in \mathbb{K} \mid k \geq s_{i0}, A_{k+1} = 0\}, \end{aligned} \tag{2.5}$$

with $t_{00} = 0$. Let $\mathbb{I}$ be the set of indices i for which s_{i0}, hence also t_{i0}, is defined. The interval $S_i = [s_{i0}, t_{i0}]$, $i \in \mathbb{I}$, is called the i-th *section*. Only the last section can be infinite. By definition s_{i0} is the first time Z_k is positive after $t_{i-1,0}$, $A_k > 0$ for all $k \in S_i$ and $A_k = 0$ between sections. If $Z_k > 0$ for at least one k, then also $A_k > 0$ for at least one k, and there is at least one section, otherwise there are none.

We further define the following, see Fig. 1:

$$\begin{aligned} Y_{i0} &= \max\{A_k \mid k \in S_i\}, \\ e_{i0} &= \min\{k \in S_i \mid A_k = Y_{i0}\}. \end{aligned} \tag{2.6}$$

The variable Y_{i0} is the maximum score obtained in section S_i and e_{i0} is the index for which it is obtained for the first time. Here and elsewhere we allow the 'maximum' to be infinite.

Recursively, define for $j > 0$,

$$\begin{aligned}
s_{ij} &= \min\{k \in S_i \mid k > e_{i,j-1}, A_k > A_{k-1}\}, \\
t_{ij} &= \min\{k \in S_i \mid k \geq s_{ij}, A_{s_{ij}-1} \geq A_{k+1}\}, \\
Y_{ij} &= \max\{A_k \mid k \in [s_{ij}, t_{ij}]\}, \qquad (2.7)\\
e_{ij} &= \min\{k \in [s_{ij}, t_{ij}] \mid A_k = Y_{ij}\}.
\end{aligned}$$

For given i, the recursion stops the first time s_{ij} is not defined.

The intervals $[s_{ij}, e_{ij}]$ are by definition non-overlapping and between any two such intervals there is at least one point, that is $s_{ij} > e_{i,j-1} + 1$, hence they cannot be adjacent. The main difference between (2.5) and (2.7) is that s_{ij} is the first time $Z_k = A_k - A_{k-1}$ is positive after $e_{i,j-1}$, whereas s_{i0} is the first time this happens after $t_{i-1,0}$.

It follows that $[s_{ij}, e_{ij}]$ is a segment (Definition 2.2): The partial sums $U_{s_{ij},k}$ are by definition positive. If $U_{k,e_{ij}}$ was non-positive for some k, the score $Y_{ij} = U_{s_{ij},e_{ij}}$ would not be maximal as required by (2.7). Hence, $[s_{ij}, e_{ij}]$ is a segment.

Definition 2.8. *The first segment in section S_i, $i \in \mathbb{I}$, is called the* independent segment *of the section and denoted by J_{i0}. The remaining segments in S_i are numbered consecutively J_{ij}, $j \in \mathbb{D}_j$ and termed the* dependent segments *of section S_i.*

The independent and dependent segments $[s_{ij}, e_{ij}]$, indicated with red and blue bars on the x-axis of Fig. 1, comprise all the maximal segments in the motivating example. This holds in general:

Proposition 2.9. *The maximal segments are precisely the segments J_{ij}, $i \in \mathbb{I}$, $j \in \{0\} \cup \mathbb{D}_i$ with score Y_{ij}. The score of a dependent segment J_{ij} depends on the score of the independent segment J_{i0} in the sense that $Y_{ij} \leq Y_{i0}$, $i \in \mathbb{I}$, $j \in \mathbb{D}_i$.*

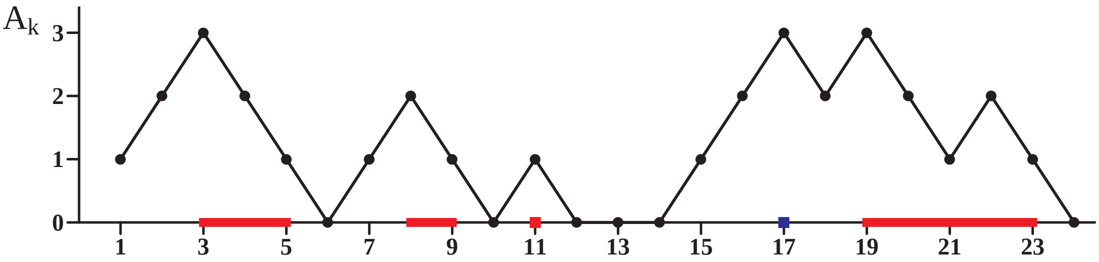
The recursions given in (2.5), (2.6) and (2.7) can be implemented using dynamical programming (details available from the authors on request). Overall the algorithm runs in time $O(|\mathbb{K}|^2)$, whereas the sections can be found in time $O(|\mathbb{K}|)$.

2.4 Reversing and extending the sequence

If $|\mathbb{K}| < \infty$, then the reversed sequence of random variables, $Z_k^r = Z_{|\mathbb{K}|-k+1}$, $k \in \mathbb{K}$, starting from the right running towards the left is well defined. By definition the maximal segments of the reversed sequence are the same as those of the forward sequence. The same is *not* true for the sequence of independent and dependent segments as they depend on the direction of the sequence, compare Fig. 2 with Fig. 1.

If the sequence Z_k, $k \in \mathbb{K}$, is extended to the right then the maximal segments do not change, potentially apart from those in the last section $S_{|\mathbb{I}|}$. If $A_{|\mathbb{K}|} \neq 0$, then the last section stops before the accumulated sum reaches zero. Adding more variables might therefore change the maximal segments of the last section. If $A_{|\mathbb{K}|} = 0$, then the maximal segments are unaffected by adding more variables. Similarly, if the sequence is extended to the left, the last section of the reversed sequence determines the segments that might change. The start and end points of the sections in the forward and the reversed sequence are not identical. Hence, we cannot identify the sections of the reversed sequence from the sections of the forward sequence (Figs. 1 and 2).

Figure 2



Landscape picture of the reverse sequence relative to Fig. 1. Here all maximal segments are independent (indicated with red bars), except for that containing position 17, which is a dependent segment and indicated by a blue coloured bar. Non-maximal segments are not shown.

3 Evaluation of Scores

We are interested in the distribution of the score Y_{ij} of a typical maximal segment. Even if the Z_ks are independent random variables, the scores will in general not be independent. Only a few theoretical results are known about the distribution of Y_{ij}. These are primarily based on random walk theory. In typically applications, however, the assumptions necessary to apply random walk theory are not fulfilled and we have to resort to other methods. We propose two simulation-based strategies.

Example 3.1. Assume as in the motivating example (Section 2.1) that a test is performed for each $k \in \mathbb{K}$ with common significance threshold $0 < \alpha < 1$. Let Z_k be 1 if the k-th test is significant and let Z_k be -1 otherwise. That is, under the null hypothesis, $Z_k = 1$ with probability α and $Z_k = -1$ with probability $1-\alpha$. The expectation $E(Z_k) = 2\alpha - 1$ is negative for $\alpha < 0.5$.

Example 3.2. Let X_k be a positive variable, for example a p-value. Define $Z_k = \log(z_\alpha / X_k)$ for some $z_\alpha > 0$. If X_k is a p-value, z_α could be the common (non-adjusted) significance level for the tests, $z_\alpha = \alpha$. If X_k is a test statistic, z_α could be the α-quantile of X_k. For example, if X_k is $\chi^2(1)$-distributed and $\alpha = 0.05$, then the α-quantile is $z_\alpha = 3.84$ and $Z_k = \log(3.84/X_k)$.

If $X_k = \alpha/e$ ($e \approx 2.7182..$) with probability α and $X_k = \alpha e$ otherwise, then we retrieve the situation in Example 3.1.

3.1 Independent Scores

Assume Z_k fulfils the condition

$$E(Z_k) < 0, \quad P(Z_k > 0) > 0, \quad \text{and} \quad \mathrm{Var}(Z_k) < \infty. \tag{3.3}$$

That is, Z_k tends to be negative but it can take positive values. It follows from the law of large numbers, that if the Z_ks are independent with common distribution, U_{nm} will eventually hit zero as m becomes large. Example 3.1 and 3.2 fulfil that $P(Z_k > 0)$ is positive. Further, $E(Z_k) < 0$ might or might not be fulfilled depending on the choice of threshold.

Let $M_0 = |\mathbb{I}|$ be the number of independent segments and $M_i = |\mathbb{D}_i|$ the number of dependent segments of section S_i.

Theorem 3.4. *Assume that Z_k, $k \in \mathbb{K}$, are independent random variables with common distribution, fulfilling condition (3.3). Assume $|\mathbb{K}| = \infty$, then the following holds:*

1. *$M_0 = \infty$ and $\sum_{i=1}^{M_0} M_i = \infty$, but M_i, $i > 0$, is finite with probability one.*
2. *The distribution of Y_{ij}, $j \geq 0$, does not depend on i.*
3. *Let $\widetilde{\mathbb{I}} \subseteq \mathbb{I}$ be a finite index set, corresponding to independent segments. The distribution of the scores factorizes as*

$$P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}}) = \prod_{i=1}^{|\widetilde{\mathbb{I}}|} P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i),$$

for $x_{ij} \in \mathbb{R}$. In particular, the scores of the independent segments form a series of independent random variables.

For any value of $|\mathbb{K}|$, we have

4. *$P(Y_{ij} \leq Y_{i0}, j \in \mathbb{D}_i) = 1$ for any $i \in \mathbb{I}$.*

The last property is general and does not require any of the assumptions of the theorem. It follows from Definition 2.8 alone.

To state the distributional results we need one further assumption. Assume in addition to the assumptions of Theorem 3.4 that there exists a number $\lambda \neq 0$, such that

$$E(e^{\lambda Z_k}) = 1 \quad \text{and} \quad E(Z_k e^{\lambda Z_k}) < \infty. \tag{3.5}$$

Hereafter let λ be such a number. If Z_k only takes a *finite* number of values then the existence of a λ fulfilling condition (3.5) follows from condition (3.3) (see Appendix).

A random variable Z is said to be a lattice variable if there is $\delta > 0$, such that Z takes values in $\mathbb{Z}_\delta = \{\delta j | j \in \mathbb{Z}\}$. Given a lattice variable with values in $\mathbb{Z}_\delta$, δ is assumed to be chosen as large as possible. If Z is not a lattice variable, then Z is said to be a non-lattice variable. Karlin and Dembo (1992) (see also Iglehart (1972)) prove that if Z_k is a lattice variable and $|\mathbb{K}| \gg i$, then for large integers y,

$$P(Y_{i0} \geq \delta y) \approx Ce^{-\delta\lambda y} \tag{3.6}$$

for some constant C. Thus, the tail distribution of Y_{i0} is approximately geometric. By applying the result to the reversed sequence we obtain an approximate distribution of the score of other independent segments. This applies in particular to the last segment, $i = M_0$, which is the first independent segment in the reversed sequence.

If Z_k is a non-lattice variable a similar result holds, however, the tail distribution of Y_{i0} is now approximately exponential (Karlin and Dembo, 1992),

$$P(Y_{i0} \geq y) \approx Ce^{-\lambda y}, \tag{3.7}$$

where y is a large real number and C a constant. The constant is characterized by the distribution of the partial sums in the lattice as well as the non-lattice case. In general, it cannot be worked out explicitly, but must be found by approximation or simulation.

Alternatively, one might restrict the observed scores of the independent segments to those bigger than a certain value y_0 for which the distributional approximation is assumed to hold. Then we obtain an approximate geometric distribution in the lattice case,

$$P(Y_{i0} \geq \delta y | Y_{i0} \geq \delta y_0) \approx (1 - e^{-\delta\lambda}) e^{-\delta\lambda(y-y_0)}, \tag{3.8}$$

and an approximate exponential distribution in the non-lattice case,

$$P(Y_{i0} \geq y | Y_{i0} \geq y_0) \approx \lambda e^{-\lambda(y-y_0)}, \tag{3.9}$$

but now without the need to determine the constant C.

Example 3.10 (Example 3.1, continued)**.** Here $\lambda = \log\left(\frac{1-\alpha}{\alpha}\right)$. The variable Z_k is a lattice variable with $\delta = 1$. It follows that the asymptotic tail distribution is $P(Y_{i0} \geq y) \approx Ce^{-\lambda y}$ for large integers y. In this case, $C \approx 1 - e^{-\lambda}$ (Karlin and Dembo, 1992), hence Y_{i0} is approximately geometric $Geo(p)$ with $p = 1 - e^{-\lambda}$.

Example 3.11 (Example 3.2, continued)**.** Recall that $Z_k = \log(z_\alpha / X_k)$. Hence λ fulfils $1 = E(e^{\lambda Z_k}) = z_\alpha^\lambda E(e^{-\lambda \log(X_k)})$. If X_k is a uniform variable (p-value), $-\log(X_k)$ is an exponential variable with intensity 1 and $z_\alpha = \alpha$. It follows that

$$\alpha^\lambda E(e^{-\lambda \log(X_k)}) = \alpha^\lambda \int_0^\infty e^{(\lambda-1)x} dx = \frac{\alpha^\lambda}{1-\lambda}.$$

Hence, there is a unique $0 < \lambda < 1$ fulfilling

$$\log(\alpha) = \frac{\log(1-\lambda)}{\lambda}, \quad \text{such that} \quad E(e^{\lambda Z_k}) = 1.$$

The tail distribution of Y_{i0} is approximately an exponential distribution $Ce^{-\lambda y}$ for some C.

3.2 Non-independent Variables

The assumption that Z_k, $k \in \mathbb{K}$, are independent random variables is very restrictive. The results in the previous section can be shown to hold also if Z_k is controlled by a Hidden Markov Model (Karlin and Dembo, 1992), which broadens the scope of applications. However, this might still be too restrictive, for example in association mapping, where the variables Z_k rarely are equally spaced along chromosomes and there might be higher order dependencies among them. In addition, only independent segments can be assigned a p-value. Here we present two bootstrap-based approaches to remedy the theoretical shortcomings.

Approach 1. Here we evaluate the score of each maximal segment against the distribution of the score of a randomly chosen maximal segment. Thus, we disregard the positional information encoded in the indices of the score Y_{ij}. Clearly, for this approach to make sense, we must require some homogeneity in distribution across the sequence. A natural requirement would be that $(Z_1, \ldots, Z_{|\mathbb{K}|})$ forms a stationary sequence, that is, the distribution of a subsequence $(Z_k, Z_{k+1}, \ldots, Z_{k+j})$ does not depend on the position k. In particular, the correlation between two variables Z_k and Z_{k+j} only depends on j.

Apply a bootstrap procedure to obtain B bootstrapped samples of the data: $(Z_1^b, \dots, Z_{|\mathbb{K}|}^b)$, $b = 1, \dots, B$. The chosen procedure will in general depend on the concrete data set. For each bootstrapped sample we find the maximal segments and let $M^b = M_0^b + \sum_{i=1}^{M_0^b} M_i^b$ denote the number of maximal segments in the b-th bootstrapped sample.

Let $\widetilde{B}$ be the number of bootstrapped samples for which $M_0^b \geq 1$ and thus $M^b \geq 1$ and let these *informative* samples be sequentially numbered $b = 1, \dots, \widetilde{B}$.

Let Y be the score of a randomly chosen maximal segment. The score Y is equal to Y_{ij} with probability $1/M$, if there are $M = M_0 + \sum_{i'=1}^{M_0} M_{i'}$ maximal segments and $M_0 \geq 1, M_i \geq j$. Hence, the distribution of Y might be approximated by,

$$P(Y \geq y) \quad \approx \quad \frac{1}{\widetilde{B}+1} \sum_{b=0}^{\widetilde{B}} \sum_{i=1}^{M_0^b} \sum_{j=1}^{M_i^b} \frac{1(Y_{ij}^b \geq y)}{M^b}, \tag{3.12}$$

where $b = 0$ denotes the original (non-bootstrapped) sample. $1(\cdot)$ is the indicator function taking the value one if the condition in parenthesis is fulfilled and otherwise it is zero. The precision of the approximation depends on the number of informative bootstrap samples $\widetilde{B}$ to the order $1/\sqrt{\widetilde{B}}$, which is a consequence of the central limit theorem.

To correct for multiple testing at level α, we might apply the correction $\alpha/E(M)$. Under reasonable assumptions, this correction controls the family-wise error (FWE), as well as the expected number of false discoveries (Type I errors; see Appendix) at level α. The expectation $E(M)$ can be approximated by the bootstrapped samples, $E(M) \approx \frac{1}{\widetilde{B}+1} \sum_{b=0}^{\widetilde{B}} M^b$. In Example 3.1 and 3.2, $E(M) \leq \alpha|\mathbb{K}|$ (α in the definition of Z_k is taken to be the same as the level at which we wish to control; the bound would be achieved if all significant tests each gave rise to a maximal segment). Hence we expect the correction factor to be at least α times smaller than the standard Bonferroni correction, $|\mathbb{K}|$.

Approach 2. Here we take the basic observation at position k to be the score of the maximal segment spanning the position. Denote this score by $Y(k)$, that is, $Y(k) = Y_{ij}$ for $k \in [s_{ij}, e_{ij}]$. If k is not in a maximal segment then $Y(k) = 0$. We seek the probability $P(Y(k) \geq y)$, where y is the observed score.

Apply a bootstrap procedure to obtain B bootstrapped samples of the data, $(Z_1^b, \dots, Z_{|\mathbb{K}|}^b)$, $b = 1, \dots, B$, and calculate the bootstrapped scores $Y^b(k)$ for each position in each bootstrapped sample. The probability $P(Y(k) \geq y)$ might be approximated by how often position k is in a bootstrapped maximal segment that has a score higher than the observed score. We obtain

$$P(Y(k) \geq y) \approx \frac{1}{B+1} \sum_{b=0}^{B} 1(Y^b(k) \geq y), \tag{3.13}$$

where $b = 0$ denotes the original (non-bootstrapped) sample.

The precision of the approximation depends on the number of bootstrapped samples B as well as the number of maximal segments obtained in each sample. In general (3.13) requires a higher B than (3.12) to obtain the same precision as we cannot use all bootstrapped segments to assess the significance of an individual segment, but only those bootstrapped segments that overlap with it. The right side of (3.13) converges as $1/\sqrt{B}$, which follows from the central limit theorem. The p-values for neighbour positions are dependent.

This procedure assigns a p-value to each position, rather than to each maximal segment, as with Approach 1. In this way the original value, Z_k (a p-value or test value), is transformed into a new p-value, which indicates the significance given to a position being in a maximal segment. If $Y(k) = 0$, then the p-value is always one and only positions in maximal segments can be declared significant. We could treat the p-values in different ways. One possibility is to declare a maximal segment significant if all positions in the segment have p-values smaller than $\alpha/E(M)$.

4 Real data example

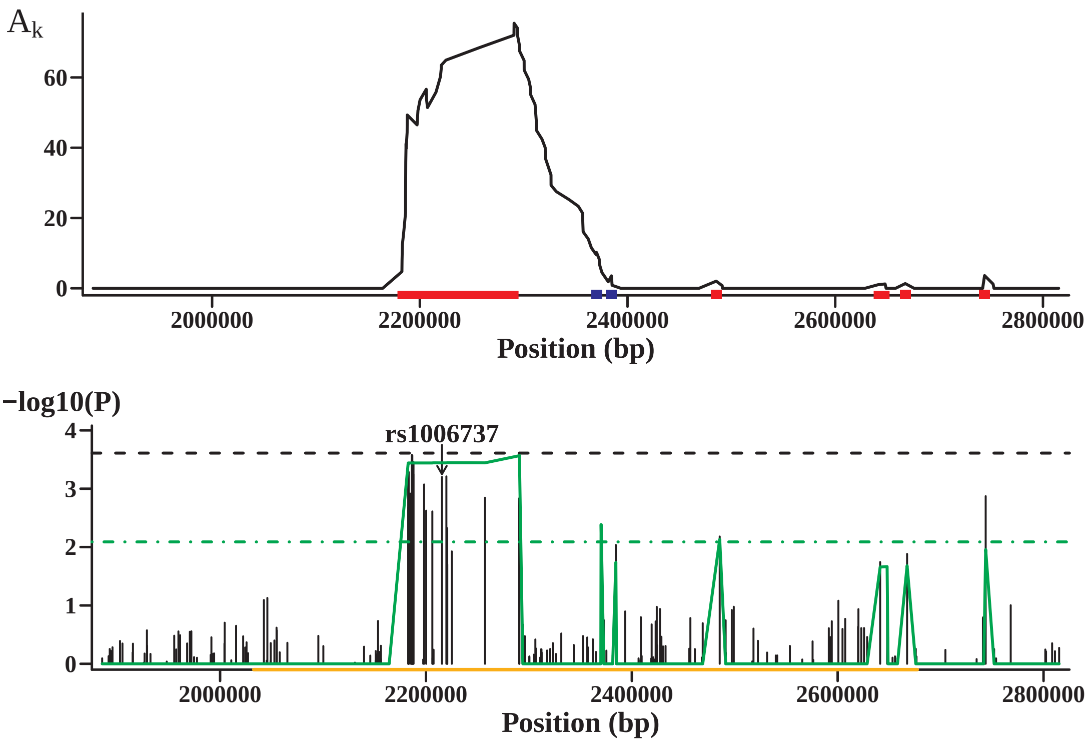
In a combined analysis of two genome-wide association studies (GWAS) by Sklar et al. (2008) signal of association with bipolar disorder was found for a single-nucleotide polymorphism (SNP), rs1006737, in the gene *CACNA1C* on chromosome 12p13.33 encoding a subunit of a calcium channel. The signal was found after combining online p-values from the WTCCC1 bipolar sample (Wellcome Trust Case Control Consortium, 2007) with p-values obtained from a combined sample of bipolar I patients from the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD) and University College London (UCL). Combined p-values were only calculated for the top 200 SNPs (out of a total of 372,193) from STEP-BD/UCL single-marker allelic association tests. In the WTCCC1 study rs1006737 was number 137,710 out of 459,446 autosomal SNPs ordering by online p-values. The involvement of *CACNA1C* in mental disorders has further been confirmed in other studies (Ferreira et al., 2008; Nyegaard et al., 2010; Ripke et al., 2013).

As a proof of concept of the Landscape method we re-examined markers in *CACNA1C* for the WTCCC1 bipolar sample (Wellcome Trust Case Control Consortium, 2007) to investigate if the association could have been identified earlier by aggregating signals. We added a buffer zone of 25% (161,175 bp) of the size of the gene region to both ends to avoid edge effects. The markers were filtered according to the description in WTCCC1 (Wellcome Trust Case Control Consortium, 2007) and genotypes with posterior probability less than 0.90 were removed (coded as missing). A total of 204 SNPs remained in this region after removing 11 monomorphic markers. Trend test p-values were calculated by logistic regression with the generalized linear model function (glm) in R. Affection status (case/control label) was shuffled and p-values recalculated 999,999 times to enable calculation of permutation-based p-values in Landscape. Fig. 3 shows the results from trend test and from using Approach 2 for non-independent variables (Section 3.2) with Z_k defined as in Example 3.2: $Z_k = \log(\frac{z_\alpha}{X_k})$ with X_k being single-marker trend test p-values and $z_\alpha = \alpha = 0.05$.

We Bonferroni corrected the threshold of significance for multiple testing by dividing the significance level α with 6.135, the mean number of maximal segments as approximated by the average number of maximal segments in the permutation-samples.

The Landscape method detects a clearly significant maximal segment around rs1006737 spanning 108 kb and consisting of 26 SNPs.

Figure 3



Upper: Landscape plot against base pair (bp) position in *CACNA1C* on chromosome 12p13.33 for $Z_k = \log(\alpha/X_k)$ where $\alpha = 0.05$ and X_k are p-values from the single-marker test shown in the lower plot. Independent and dependent segments are indicated on the x-axis with red and blue bars, respectively. **Lower:** Results from single-marker trend tests, none of which were significant at level $\alpha = 0.05$ after Bonferroni correction for the 204 tests (threshold indicated with the black dashed line). The green line is from using Approach 2 of Landscape with Z_k as above and using 999,999 permutation-based p-values. Bonferroni corrected threshold adjusted for the mean number of maximal segments ($E(M) = 6.135$) is indicated with the green dash-dotted line. The orange line on the x-axis indicates the gene region of *CACNA1C*.

5 Discussion

We have developed a method to aggregate sequentially ordered statistics and provided different means to assess the significance of the aggregated scores, that is, the scores of maximal segments. If the original variables (Z_k) are dependent, the aggregated scores will in general also be dependent. As shown using a real data set, the aggregated score might be significant without the individual p-values being significant. Thus, our method may be a useful supplement to standard procedures relying on evaluation of test statistics individually.

In a sense, one can consider the aggregated score a smoothening of the individual test statistics (or p-values) as the values are ‘smoothed’ with the values of the surrounding positions. The smoothed value of position k is the score $Y(k)$ of the maximal segment containing that position. It might be written as

$$Y(k) = \max\{U_{nm} | U_{nk'} > 0, U_{k'm} > 0, \forall k' \in [n,m], n \le k \le m\}.$$

Our method assigns p-values to maximal segments or positions. These we have to correct for multiple testing but now each of them borrow from their neighbours in contrast to the original p-values that are based on individual tests. Here we have applied a simple Bonferroni procedure for multiple testing but more sophisticated techniques could likewise be used.

Finally, one might be able to learn the parameter λ in (3.6) and (3.7) from a genome-wide empirical distribution, thereby extending the realm of application to situations with structural dependencies.

Acknowledgements

The study was supported by grants from the Danish Strategic Research Council (2101-07-0059) (GEMS consortium), the Lundbeck Foundation, Denmark, and the Danish Cancer Society. This study makes use of data generated by the Wellcome Trust Case-Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the WTCCC data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475.

Appendix

Proof of Proposition 2.9. Consider an interval $[s,t] = [s_{ij},t_{ij}]$ for some i,j. We have already shown that $[s,t]$ is a segment. It is also maximal: to show this, we must prove that there is no other segment $[n,m]$ containing it. First note that $Z_k \leq 0$ for all k between two intervals found by the algorithm, that is, for all k such that $t_{i,j-1} < k < s_{ij}$ or $t_{i-1,j} < k < s_{i0}$ for some i,j (in the latter $t_{i-1,j}$ refers to the last interval in the previous section S_{i-1}). Thus, $[n,m]$ cannot start or end between intervals as this would imply that Z_n or Z_m is non-positive. A segment $[n,m]$ cannot bridge two intervals either. If this was so, we would have $U_{nk} > 0$ for $n \leq t < k < s'$, where $[s,t]$ and $[s',t']$ are two intervals. But

$$U_{nk} = U_{n,t+1} + U_{t+2,k} = (A_t + Z_{t+1}) - A_{n-1} + U_{t+2,k}.$$

The first and last terms are non-positive by definition. Hence $U_{nk} \leq 0$, a contradiction, and any interval is a maximal segment.

To prove the reverse, we note that the intervals are disjoint. Hence, if there are more maximal segments than the intervals found by the algorithm they must be between intervals. However, as noted above, Z_k is non-positive between intervals, hence there cannot be more maximal segments. □

Proof of Theorem 3.4. Assume $|\mathbb{K}| = \infty$. (1) It follows from the law of large numbers that $\frac{1}{m-n+1}U_{nm} \to E(Z_n) < 0$ as $m \to \infty$. Hence the partial sum will eventually become negative with certainty. Assume there is only a finite number of independent segments M_0 and let s_{M_0} be the start of the last. Then either (i) $U_{s_{M_0},s_{M_0}+n-1} > 0$ for all n or (ii)

$U_{s_{M_0},s_{M_0}+n-1} \leq 0$ for some n and $Z_{s_{M_0}+m} \leq 0$ for all $m > n$. In the first case,

$$\begin{aligned} & P(U_{s_{M_0},s_{M_0}+n-1} > 0, n \geq 1) \\ = & \sum_{m=1}^{\infty} P(U_{s_m,s_m+n-1} > 0, n \geq 1 \mid M_0 = m) P(M_0 = m) \\ = & \sum_{m=1}^{\infty} P(U_{1n} > 0, n \geq 1) P(M_0 = m) \\ = & P(U_{1n} > 0, n \in \mathbb{K}) > 0, \end{aligned} \tag{A.1}$$

where the second equality follows from Z_k being independent and identically distributed (i.i.d.). This contradicts that the partial sum eventually becomes negative, hence $M_0 = \infty$. As for the second case, $P(Z_k > 0) > 0$, by assumption. Hence the probability that all $Z_{s_{M_0}+n}$ are non-positive is zero and we conclude again $M_0 = \infty$. For each independent segment, there is a positive probability of a dependent segment. Hence, the number of dependent segments will be infinite.

(2) It follows similarly to (1) by conditioning on the start of the maximal segment and using that the Z_ks are i.i.d.

(3) Is proven by induction on the size k of $\widetilde{\mathbb{I}}$. For $k = 1$, the claim is obviously true. Assume it is true for some $k \geq 1$. The probability $P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}})$ is

$$\sum_{e=1}^{\infty} P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}} \setminus \{1\} \mid A_1) P(A_1),$$

where $A_1 = \{T_1 = e, Y_{1j} \leq x_{1j}, j \in \mathbb{D}_1\}$ and T_1 is the end of the first section. Since Z_n are independent variables and $|\mathbb{K}| = \infty$, the first probability in the sum is independent of A_1 (the sequence $Z_{T_1+1}, \ldots,$ has the same distribution as $Z_1, \ldots$). Hence, the probability is

$$\begin{aligned} & \sum_{e=1}^{\infty} P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}} \setminus \{1\}) P(A_1) \\ & \qquad = P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}} \setminus \{1\}) \sum_{e=1}^{\infty} P(A_1) \\ & \qquad\qquad = P(Y_{ij} \leq x_{ij}, j \in \mathbb{D}_i, i \in \widetilde{\mathbb{I}} \setminus \{1\}) P(Y_{1j} \leq x_{1j}, j \in \mathbb{D}_1). \end{aligned}$$

The claim now follows from the induction hypothesis. □

Existence of λ. Assume Z_k fulfils condition (3.3) and that Z_k only takes a finite number of values. Hence $f(\lambda) = E(e^{\lambda Z_k})$ is finite for all $\lambda \in \mathbb{R}$. The derivative of $f(\lambda)$ is $f'(\lambda) = E(Z_k e^{\lambda Z_k})$ for all λ (Hoffmann-Jørgensen, 1994). Hence $f(0) = 1, f'(0) < 0$, since $E(Z_k) < 0$, and $f(\lambda) \to \infty$ as $\lambda \to \infty$, since $P(Z_k > 0) > 0$. Further, $f(\lambda)$ is convex for all λ (Hoffmann-Jørgensen, 1994). Then it must be that there is a unique $\lambda_0 > 0$ such that $f(\lambda_0) = E(e^{\lambda_0 Z_k}) = 1$. □

Approach 1: Type I errors. Given α, choose y_α, such that $P(Y \geq y_\alpha) \leq \alpha/E(M)$. We have

$$P(Y \geq y_\alpha) = \sum_{m=1}^{\infty} P(Y \geq y_\alpha|m)P(m), \tag{A.2}$$

where "m" is short for the event $\{M = m\}$. If M and $U(M) = P(Y \geq y_\alpha|M)$ are negatively correlated variables, then the FWE and the expected number of Type I errors are controlled at level α. The assumption is plausible as more maximal segments should reduce the general score of them.

It can be proven in the following way. If there are m maximal segments, list them sequentially without regards to whether they are dependent or independent segments. The expected number of Type I errors is

$$E\left(\sum_{i=1}^{M} I_j\right) = \sum_{m=1}^{\infty}\sum_{j=1}^{m} E(I_j|m)P(m), \tag{A.3}$$

where $I_j = 1$ if the score of the jth maximal segment is $\geq y_\alpha$ and otherwise $I_j = 0$. Since Y is the score of a randomly chosen maximal segment, then

$$U(m) = P(Y \geq y_\alpha|m) = \frac{1}{m}\sum_{j=1}^{m} E(I_j|m).$$

Hence, from (A.3),

$$E\left(\sum_{i=1}^{M} I_j\right) = \sum_{m=1}^{\infty} mU(m)P(m).$$

If M and $U(M)$ are negatively correlated variables,

$$E\left(\sum_{i=1}^{M} I_j\right) \leq E(M)\sum_{m=1}^{\infty} U(m)P(m) = E(M)P(Y \geq y_\alpha),$$

where the equality follows from (A.2). Recall that y_α is chosen such that $P(Y \geq y_\alpha) \leq \alpha/E(M)$, hence $E\left(\sum_{i=1}^{M} I_j\right) \leq \alpha$ and the expected number of Type I errors is controlled at level α. Since FWE is less than the expected number of Type I errors, we also have control of FWE at level α. □

References

Cheverud, J. M. (2001). "A simple correction for multiple comparisons in interval mapping genome scans." *Heredity* **87**: 52–58.

Ferreira, M. et al. (2008). "Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder". *Nat. Genet.* **40**: 1056–1058.

Fisher, R. A. (1932). *Statistical methods for research workers*. Oliver and Boyd.

Hoffmann-Jørgensen, J. (1994). *Probability with a view toward statistics*. Vol. I. New York: Chapman & Hall.

Iglehart, E. (1972). "Extreme values in GIG1 queue". *Ann. Math. Stat.* **43**: 627–635.

Karlin, S. and Altschul, S. F. (1990). "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." *Proc. Natl. Acad. Sci. U S A* **87**: 2264–2268.

Karlin, S. and Dembo, A. (1992). "Limit Distributions of Maximal Segmental Score among Markov-Dependent Partial Sums". *Adv. Appl. Prob.* **24**: 113–140.

Nyegaard, M. et al. (2010). "CACNA1C (rs1006737) is associated with schizophrenia". *Mol.Psychiatry* **15**: 119–121.

Nyholt, D. R. (2004). "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other." *Am. J. Hum. Genet.* **74**: 765–769.

Ripke, S. et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia". *Nat. Genet.* **45**: 1150–1159.

Simes, R. J. (1986). "An Improved Bonferroni Procedure for Multiple Tests of Significance". *Biometrika* **73**: 751–754.

Sklar, P. et al. (2008). "Whole-genome association study of bipolar disorder". *Mol. Psychiatry* **13**: 558–569.

Stouffer, S. A. et al. (1949). *The American soldier, vol 1. Adjustment during army life*. Princeton: Princeton University Press.

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". *Nature* **447**: 661–678.

Appendix A

A few more technical details

A.1 A bit more on logic regression

A.1.1 The origin of *logic regression* and misuses of the term

The adaptive regression methodology called *logic regression* was introduced by Ruczinski (2000) in his PhD thesis[36], of which Ruczinski et al. (2003) appears to be a condensed version. Apparently Ruczinski, Kooperberg and LeBlanc also wrote a technical report from Fred Hutchinson Cancer Research Center in 2001 (c.f. Kooperberg et al., 2001) about logic regression but we have not been able to track this down. Searching (October 5, 2011) for the term ”logic regression” in PubMed (www.ncbi.nlm.nih.gov/pubmed/) reveals that this term was mentioned in a few earlier publications, the first time being in Troeng et al. (1994) which predates Ruczinski's PhD by more than five years. However, in Troeng et al. (1994) it simply appears to be a typo in the abstract, as the methods used in the paper are logistic regression and Cox regression. Another mention is in a Russian journal Voprosy Onkologii (Problems in Oncology) 1996;42(1):48-52, where *multifactorial logic regression* was used according to the translated abstract. However, a search on Google reveals no other matches on the term ”multifactorial logic regression” than this paper whereas a search for ”multifactorial logistic regression” returns thousands of hits so this was probably a typo or erroneous translation. The term ”logic regression” is also mentioned in the (translated) abstract of a paper in another Russian journal, Zh Nevrol Psikhiatr Im S S Korsakova (Zhurnal nevrologii i psikhiatrii imeni S.S. Korsakova / Ministerstvo zdravookhraneniia i meditsinskoi promyshlennosti Rossiiskoi Federatsii, Vserossiiskoe obshchestvo nevrologov [i] Vserossiiskoe obshchestvo psikhiatrov) 1999;99(5):32-40, but this seems to be (a follow-up of?) the same study as in Acta Neurol Scand 1996: 94: 386-394 (which is in English!), where again they use logistic regression and do not mention logic regression. The last paper reporting use of logic regression and appearing around the same time as Ruczinski (2000) is a Spanish paper in Rev Esp Salud Publica (Revista española de salud pública) 2001;75(1):81-8, where the sentence ”Logic Regression was employed for calculating the odds ratio adjusted by age, sex and by the intake of foods and wine” can be found in the (translated) abstract. It is quite likely again either wrongly translated, a typo or an erroneous use the word ”logic” where instead ”logistic” should have been used.

As a curiosity, 14 more abstracts from this same Spanish journal contained *logic regression* where it appears from the context (stating odds ratio e.g.) that the correct term would

[36] http://kooperberg.fhcrc.org/logic/documents/ingophd-logic.pdf

have been *logistic regression*: Rev Esp Salud Publica 2001;75(6):529-39, 2002;76(6):673-82, 2003;77(1):143-50, 2003;77(2):287-95, 2004;78(3):367-77, 2004;78(4):481-92, 2004;78(4):527-37, 2005;79(1):47-57, 2005;79(1):59-67, 2005;79(4):465-73, 2005;79(5):541-9, 2005;79(5):559-67, 2006;80(4):335-47, 2008;82(3):315-22. Also one French publication (Sante Publique 2005;17(2):265-80) erroneously translated ”régression logistique” to ”logic regression”[37] and further two French publications states ”logic regression” where it should have been ”logistic regression”: Sante Publique 2007;19(6):489-97 and Rev Laryngol Otol Rhinol (Bord) 2010;131(4-5):247-51. In essence none of the non-English papers containing ”logic regression” in the abstract had used this term correctly. Moreover, logistic regression were also misprinted (misunderstood or badly translated) as logic regression in the abstracts of the following English-language papers: Disabil Rehabil Assist Technol 2010;5(5):318-22, Soc Sci Med 2009;68(4):643-53, BMC Infect Dis 2007;7:75, BMC Infect Dis 2006;6:113, Lancet 2006;368(9530):130-8, and J Asthma 2005;42(10):833-7. The worst of these were the BMC Infect Dis 2006 paper, where the wrong word was used also in the methods section. Finally, in Oral Oncol. 2011;47(7):588-93 they do use logic combinations of dichotomous variables but in quite another context than that of logic regression as defined by Ruczinski et al. (2003).

In conclusion, we find it evident that *logic regression* was introduced and coined by Ruczinski (2000) as also noted in Ruczinski et al. (2003).

A.1.2 Implementations

Originally, the logic regression approach by Ruczinski et al. (2003) was implemented as a stand-alone program xlogic in Fortran 90 (version 0.1.3 dates back to July 17, 2001) but at least since January 2003 (version 1.1.1 is from January 31, 2003) it has been implemented as the R package LogicReg though still with a core written in Fortran which is then called by R. A greedy search algorithm and a Markov chain Monte Carlo (MCMC) sampler was implemented with version 1.3.0 January 15, 2004 (Kooperberg et al., 2005). The following web site contains various sources of information and links to the software: `http://kooperberg.fhcrc.org/logic/`. The current version of LogicReg (1.5.5, December 3, 2013) handles classification, linear regression, logistic regression, proportional hazards model (Cox regression), exponential survival model, and by example of ”writing your own scoring function”, conditional logistic regression. Greedy search algorithm, simulated annealing and MC logic regression are implemented together with methods to do cross-validation and permutation-based tests for model selection.

A.1.3 Terminology of logic regression

The term *Boolean expression* (logic expression or Boolean logic expression) covers any Boolean combination of binary variables, where the Boolean combinations used are logical AND ($\wedge$), OR ($\vee$) and NOT (^{c}). Let us denote the binary predictors $x_1, x_2, \ldots, x_k$; then an example of a Boolean expression could be $L_j = (x_2 \vee x_4^c) \wedge x_7$ or by use of words: if {(2nd predictor is present OR 4th predictor NOT is present)} AND {7th predictor is present} then L_j is true, i.e. the state is predicted. The Boolean domain {true,false} is often represented by {1,0} and it may come in handy to think of the AND operation as a kind of multiplication with 0 or 1, and the OR as an addition (though some rules for these operators are not the same as for multiplication and addition) where the result is true if doing so returns something greater than zero, and false if

[37] There may be an explanation to this as *logistique* exists both as the adjective *logistic* and as the substantive (noun) *mathematical logic*

it is zero. As examples: {true AND false} returns false (1*0=0), {true OR false} returns true (1+0=1) and so fort. In other words x_i and x_i^c are really the indicator variables $\mathbb{1}_{[x_i=1]}$ and $\mathbb{1}_{[x_i=0]}$, respectively.

To enable coding of SNP genotypes, we need two variables for each locus. Let S_i denote a SNP and let S_{iD} be an indicator of a genotype which is <u>not</u> of the homozygous reference type (true if there is at least one variant/minor allele), and let correspondingly S_{iR} indicate the homozygous variant type (i.e. 1 if both chromosomes contain the minor allele of S_i). Note that the homozygous reference type can be identified by S_{iD}^c. The use of Boolean expressions now enables us to define models with higher order interactions between genotypes, e.g. (as in Schwender et al., 2010) $L = (S_{1D}^c \wedge S_{2R}) \vee (S_{3D} \wedge S_{4D})$ or in words: IF S_1 is of the homozygous reference genotype AND S_2 is of the homozygous variant genotype OR both S_3 and S_4 are NOT of the homozygous reference genotype, THEN the subject has an elevated risk. Note that if S_{iD} is present in the final logic tree then a dominant model fits the data best for SNP S_i whereas a recessive model is better if S_{iR} enters. The adaptive search algorithm will remove redundant statements so S_{iD} and S_{iR} will not be present in the same logic tree because $S_{iD} \wedge S_{iR} \equiv S_{iR}$ and $S_{iD} \vee S_{iR} \equiv S_{iD}$ (Kooperberg et al., 2001). Methods to remove redundancy from the models were treated by Ruczinski (2000). Inclusion of more than one logic tree enables the modeling of additive, multiplicative and co-dominant genetic models, and in these multiple tree variants it may occur that both S_{iD} and S_{iR} are present, though still not in the same logic tree.

The same Boolean expression may be represented in many different ways, and a special case is the so-called *Disjunctive Normal Form* (DNF) which is an OR-combination of AND-combinations. As an example $L = (S_{1D}^c \wedge S_{2R}) \vee (S_{3D} \wedge S_{4D})$ is a DNF but could also have been represented by e.g. $L = (S_{1D}^c \vee S_{3D}) \wedge (S_{1D}^c \vee S_{4D}) \wedge (S_{2R} \vee S_{3D}) \wedge (S_{2R} \vee S_{4D})$. Note that in a DNF L is true if at least one of the AND-combinations return true. The advantage of DNFs is that interactions can be directly identified as the AND-combinations (Schwender et al., 2008). So in the example we have two interactions. Furthermore, Schwender et al. (2008) noted that only minimal AND-combinations (denoted *prime implicants*) should be included in DNFs, that is redundant (conjugate) letters should be omitted. Algorithms to find the prime implicants and DNFs were treated in Schwender (2007) and included in logicFS (Schwender, 2007; Schwender et al., 2008). In conclusion, the use of DNFs simply makes interpretation of the expression easier.

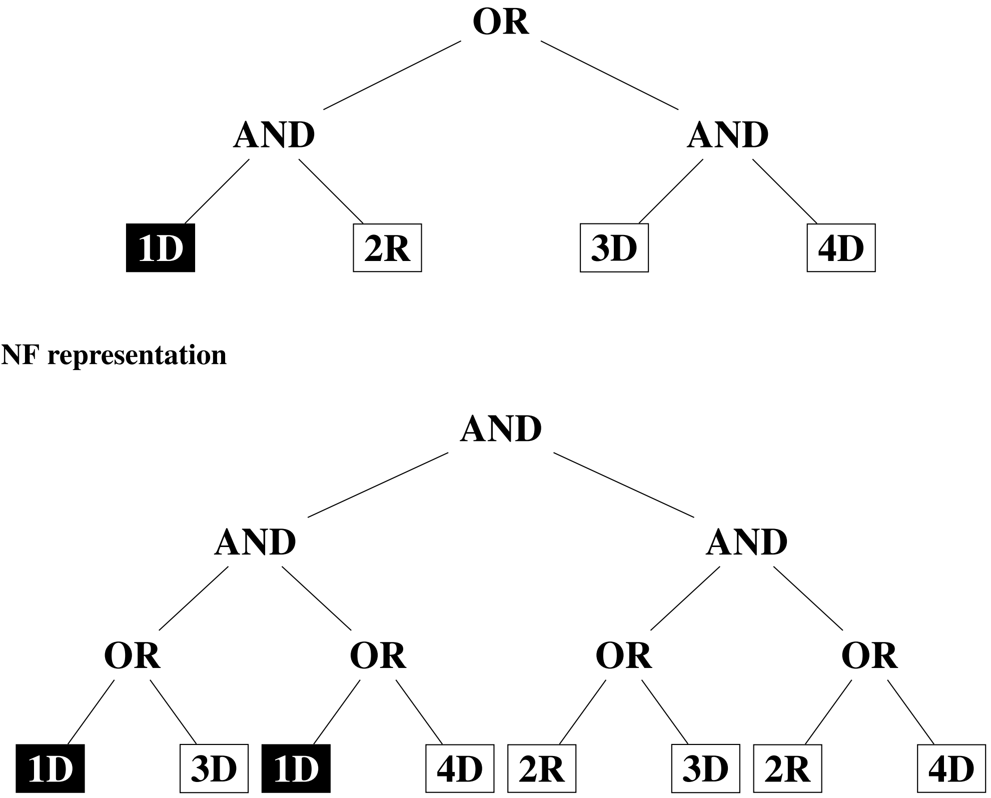
Another and equivalent way to represent a Boolean expression is by use of a so-called *logic tree*. In logic trees each knot have either zero or two children and leaves are always (conjugate) letters while knots that are not leaves always is either an AND or an OR statement. We refer to Ruczinski (2000) for further discussion on various representations and simplifications of Boolean expressions and the relationship and differences between logic trees and decision trees. The two representations of L, drawn as trees, are shown in figure A.1. Here for brevity we have left out the S's and white letters on black background indicate conjugate e.g. S_{1D}^c.

A.1.4 The regression framework

The regression framework is maybe the most appealing feature of logic regression. In essence it is a generalised linear model

$$g(E[Y]) = \beta_0 + \sum_{t=1}^{T} \beta_t L_t,$$

where g is a link function and L are logic expressions (trees) consisting of Boolean combinations of binary covariates. The models $g(E[Y])$ are referred to as *logic models* (c.f. Ruczinski et al.,



(a) DNF representation

(b) Another representation

Figure A.1 Two representations of a logic tree example
(a) DNF representation and **(b)** another possible representation of the same logic tree.

2003). This framework can be used for many different types of outcome (continuous, categorical, counts, time-to-event) with appropriate link functions as long a score function can be defined to reflect the quality (or fit) of the models considered, i.e. as long as we have a score measure that can be used in the annealing algorithm (see appendix A.1.5). During this search for the best combination of boolean expressions, i.e. the combination minimising the score, the parameters $\beta_0, \beta_1, \ldots, \beta_t$ of the model are estimated simultaneously.

Two important examples of responses and score functions are given here:

1. If the response is continuous and can be assumed to stem from a normal distribution, i.e. we consider a usual general linear normal model (linear regression), then the residual sums of square
$$RSS = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$
can be chosen as score function.

2. If the response is binary (e.g. affected/unaffected in case-control studies) then logistic regression is often used, i.e. a binomial distribution with the logit link function
$$g(E[Y]) = g(\pi) = \log(\frac{\pi}{1-\pi}).$$

Here the score function used in LogicReg is the deviance function

$$D = -2\log\left(\frac{\text{likelihood of the model to be assessed}}{\text{likelihood of the saturated model}}\right)$$

The model can be extended by inclusion of other covariates such as gender, age and socioeconomic factors (Schwender et al., 2010):

$$g(E[Y]) = \beta_0 + \sum_{c=1}^{C} \gamma_c X_c + \sum_{t=1}^{T} \beta_t L_t.$$

These covariates, $X_1, \ldots, X_C$, may be continuous, discrete, categorical or any other type we could think of to include in a regression, i.e. they do not necessarily have to be dichotomous. The parameters for these covariates will be estimated but do not enter the annealing search. For a schematic view of the algorithm see Algorithm 2 in Schwender et al. (2010).

Another way of allowing such covariates is to dichotomise them either by a pre-determined level/limit or via the simulated annealing search by defining a new move that changes this level, see Ruczinski (2000). Categorical variables can be handled similarly or may be coded by use of binary dummy variables (1 if a person belong to the category and 0 otherwise), one dummy variable for each level except one—the last of the levels will be captured in the overall mean, i.e. the parameter β_0.

A.1.5 Searching logic tress with simulated annealing

In the machine learning method *logic regression* a set of k binary predictors are used to search for gene-gene and gene-environment interactions in case-control studies. The method searches over the space of 2^k different combinations each of which can predict a zero or one (control or case, say), that is in principle 2^{2^k} possible combinations also referred to as *prediction scenarios* (see Ruczinski, 2000). The number of combinations therefore grows with double exponential speed and become incomprehensible large even for a relatively small number of predictors. Consequently a *simulated annealing* search algorithm is used to find the *best fitting* model.

Moving between logic trees

For a set of predictors, a given logic tree can in principle be reached from any other logic tree (of the predictors) in a finite number of moves given by deleting a leaf, splitting a leaf (using AND or OR and another predictor/leaf), alternating a leaf (using another predictor), or alternating an operator (AND instead of OR, or vise versa). Another two moves are defined as they enhance the performance of the algorithm (Ruczinski et al., 2003): pruning a branch (i.e. removing a branch) and growing a branch (i.e. adding another branch). We have depicted the moves in figure A.2 and they are described more thorough in many of the logic regression papers (e.g. Ruczinski, 2000; Kooperberg et al., 2001; Ruczinski et al., 2003; Schwender et al., 2008). So there are six permissible moves in the process of growing a tree. The two *alternating* moves are their own countermove, while the remaining four are move/countermove pairs. It is crucial that we can get back and forth to avoid breakdown of the Markov chain theory behind the simulating annealing. If no limits are imposed on model size, the underlying Markov chain will be irreducible as it is then possible to get between any two states within a finite number of steps. Furthermore, the Markov chain is aperiodic and therefore also ergodic, which in principle ensures convergence of the search algorithm.

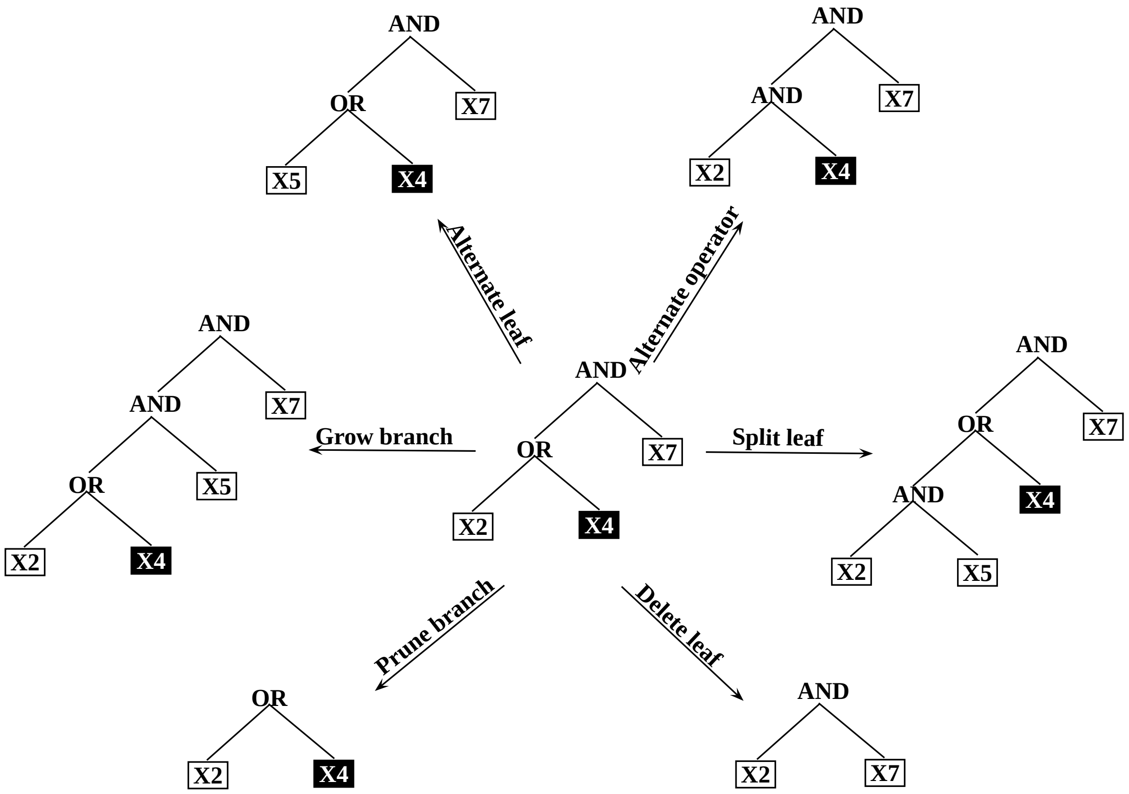


Figure A.2 Permissible moves in a simulated annealing search of logic trees
There are six permissible moves in the process of growing a tree in the simulated annealing search for the *best fitting* model (logic tree). For each move there is a counter move to ensure convergence. The two *alternating* moves (alternate leaf and alternate operator) are their own countermove, while the remaining four are move/countermove pairs.

The model search can be extended by allowing new trees to be added (starting with one leaf only), or present trees (with one leaf) to be removed. The addition of new trees can only be done though until the chosen upper limit on the number of trees has been reached. Correspondingly, the upper limit on the number of leaves in each tree constraints the splitting leaf and growing branch moves. These limitations break the ergodicity of the Markov chains and thus there is no guarantee that a global optimum will be found.

Simulated annealing

As noted by Ruczinski et al. (2003), simulated annealing is not a new development. It origins back to Kirkpatrick et al. (1983) and Cerny (1985) as an adaptation of an even older method, the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). The process is defined on some state space $\mathscr{S}$ which depends on the restrictions. If we pose the restriction of using exactly one tree in the model then $\mathscr{S}$ is the set of all possible logic trees of the predictors (plus the empty ”tree”). If any number of trees are allowed the state space (extending the set of moves correspondingly) consists of all possible combination of trees as well. For the sake of simplicity let us consider the one tree only model. Each state $s \in \mathscr{S}$ is thus a tree and two states $s, \tilde{s} \in \mathscr{S}$ are said to be adjacent (or neighbors) if only a single move (from the permissible set of moves) is needed to get from one state to the other. We let $\mathscr{M}$ denote the subspace in $\mathscr{S} \times \mathscr{S}$ defined by the adjacent pairs, i.e. $(s,\tilde{s}) \in \mathscr{M} \subseteq \mathscr{S} \times \mathscr{S}$. The elements of $\mathscr{M}$ are called moves. If k moves are needed to get from state s to state $\tilde{s}$ then $(s,\tilde{s}) \in \mathscr{M}^k$ are said to be connected via a set of k moves. The state space is assumed to be finite. To compare the logic models (or just trees) we need an objective function referred to as the score function $\omega : \mathscr{S} \to \mathbb{R}$ which quantifies the *quality* (fit) of each state by its score. We assume without loss of generality that small is better and since we assume the state space to be finite the existence of (at least one) minimal score ω_0. If classification is the goal of using logic regression there is usually only one tree and the score function is taken to be the misclassification rates of the logic tree, see Algorithm 1 in Schwender et al. (2010). More details on the use of logic regression for classification can be found in e.g. Ruczinski et al. (2003).

Now basically, from the present state $s \in \mathscr{S}$ the simulating annealing procedure picks a move $m \in \mathscr{M}$ (by some selection scheme) and thereby propose a new state $\tilde{s}$. The scores $\omega = \omega(s)$ and $\tilde{\omega} = \omega(\tilde{s})$ are compared and the proposal is always chosen if it has a better score and it is chosen with some acceptance probability if the score is worse. The acceptance probability is given by an acceptance function $\zeta : {\mathbb{R}_+}^3 \to (0,1]$ which assigns a probability to a pair of scores and some positive real number, the temperature parameter τ, which reflects for how long time the annealing chain has run. The influence of the temperature on the acceptance probability is such that the probability of accepting a worse score becomes smaller as we run through the annealing scheme and converges towards zero. So in the beginning the state space is searched more broadly while later being narrowed down to the (hopefully) right neighborhood. This way the risk of ending in a local (but not global) optimum is lowered as it is possible to jump to a better scoring state, which is connected to the present state via $k > 1$ moves, even when all adjacent states to the present state has a worse score. In Ruczinski (2000) various considerations involving Markov chain properties leads to the following acceptance function

$$\zeta(\omega,\tilde{\omega},\tau) = \min(1, \exp(-\frac{\tilde{\omega}-\omega}{\tau}))$$

which has also often been used in the literature (Ruczinski, 2000). The temperature τ is determined by a so-called cooling scheme. The cooling scheme should be such that less time is spend in the beginning, where almost all moves are accepted, and towards the end of the scheme, where virtually all moves are rejected. Thus most of the time in the algorithm should be spend somewhere in the middle, a period which Ruczinski et al. (2003) refers to as the ”crunch time”, i.e. the period where hopefully the right neighborhood containing the global optimum is chosen by the algorithm.

In Ruczinski et al. (2003) they keep the temperature fixed a period of time until a predetermined threshold number of accepted moves are reached. Then the temperature is lowered (typically in equal decrements on a log10 scale) and a new sequence of homogeneous (constant temperature) Markov chains are run until the threshold is reached and so fort. Apart from this, they also set a maximum number of iterations for each chain, after which the temperature is lowered (if the above mentioned threshold was not reached first). The threshold will only be reached in the beginning of the annealing process while later all chains are run in full length as determined by the ”maximum number of iterations” parameter. A threshold of 1-10% of the intended number of iterations was typically used by Ruczinski et al. (2003). Furthermore a start and a stop temperature needs to be chosen before running the simulating annealing. Instead of a cleverly chosen lower temperature there may be a criterion of when to stop the process, such as no improvement of the score (no moves accepted) for a considerable number of consecutive chains (temperatures)—in practice 10-20 chains (c.f. Ruczinski et al., 2003). Speed may also be gained by keeping track of the scores from states (trees) that has already been visited. By doing this the computation of the scores need only to be done once for each state. This is especially cost-saving in the last part of the run (at low temperatures) where the same state may be proposed multiple times.

Making the decisions on the various parameters involves a bit of trial and error as the optimal choice depends on the data. The highest and lowest temperature values are not that important if the criterions mentioned above are applied. The starting value (highest temperature) should simply be large enough that we more or less do a random walk in the beginning, and the lowest temperature should be so small that the procedure is stopped because no further improvement of the score is observed. But of cause, there is no reason to choose the starting temperature too far out. The acceptance rate should therefore be monitored and a lower starting temperature chosen if it takes too many chains to reach the ”crunch period”. The chain needs to be run long enough that the chain is close to its limiting distribution, i.e. close to stationarity. In Ruczinski et al. (2003) they mention 10,000 - 100,000 as the chain length they have been using. When doing the first trials to set the various parameters of the algorithm, it may be an idea to use fairly short chains (in the order of 10,000’s, say) while later, longer chains (100,000 or more) should be used too increase precision/convergence. The reason for choosing shorter chains is merely to decrease computation time. Also, the temperature step size (or equivalently the number of chains between two subsequent powers of 10) have to be chosen. According to Ruczinski et al. (2003) the temperature is usually decreased by a factor between 0.91 and 0.98 ($10^{-1/25}$-$10^{-1/100}$), corresponding to 25-100 chains between the powers of 10. In practice, numbers z_{start}, z_{end} and n_{iter} can be input to the software to define a temperature cooling scheme going from $10^{z_{start}}$ to $10^{z_{end}}$ decreasing each step by a factor of $10^{-1/n_{iter}}$.

In practice some limits on tree size (the number of leaves) and number of trees, t, are also needed. One reason is interpretation of the resulting model while another is computational due to the fact that Ruczinski et al. (2003) fit all trees in the model simultaneously. The risk of over-fitting may be a third reason. According to Ruczinski, they usually set the limit at a maximum of 8-16 leaves per tree and a maximum of five trees (if for more than one), and case studies showed that 1-3 trees were optimal. But of course these numbers depend on the number of predictors and on how large models we are willing to interpret, so no fixed rules can be given: ”Unfortunately, this is more of an art than a science, and there are no hard and fast rules how to find the best possible annealing algorithm.” (c.f. Schwender et al., 2010).

A.2 Landscape R scripts

The R scripts used for calculation of Landscape measures in the motivating example. First source the code of the *Landscape.fct* from the next three pages in R. Then the following lines of code utilises this function:

```
# --- Reading data for motivating example: ---
Z.k <- c(-1,-1,1,1,1,-1,-1,1,1,-1,1,-1,-1,-1,-1,-1,1,-1,1,1)
K <- length(Z.k); k.idx <- 1:K
# --- Doing calculations with Landscape.fct: ---
res <- Landscape.fct(Z.k)
# --- Calculate independent segments: ---
N.seg <- length(res$sij); ind.seg <- matrix(NA,ncol=2,nrow=N.seg)
for(i in 1:N.seg){ind.seg[i,] <- c(res$sij[[i]][1],res$eij[[i]][1])}
# --- Calculate dependent segments: ---
N.dep.seg <- sapply(1:N.seg,FUN=function(x){length(res$sij[[x]])-1})
dep.seg <- matrix(NA,ncol=2,nrow=sum(N.dep.seg))
n <- 0
for(i in 1:N.seg){
    if(N.dep.seg[i]>0){
        for(j in 2:(N.dep.seg[i]+1)){
            n <- n+1
            dep.seg[n,] <- c(res$sij[[i]][j],res$eij[[i]][j])}}}
# --- Show sections: ---
sec <- res$sect
cat("Maximal segments:",paste("[",sec[,1],",",sec[,2],"]"))
# --- Show maximum segments: ---
max.seg <- res$max.segm
cat("Maximal segments:",paste("[",max.seg[,1],",",max.seg[,2],"]"))
# --- Plot A_k: ---
A.k <- res$A.k
plot(k.idx,A.k[-c(1,(K+2))],type="b",xlab="Position",ylab="A_k")
# --- Find permutation-based p-values for Y(k) ---
# observed values:
Yk.obs <- res$Y.k #Yk.obs <- Landscape.fct(Z.k,report="Y")
n.perm <- 99; Y.perm <- matrix(0,ncol=length(Z.k),nrow=n.perm)
set.seed(11062014)
# permutation values:
    for(i in 1:n.perm){
    z.p <- sample(c(-1,1),size=K,replace=TRUE)
    Y.perm[i,] <- Landscape.fct(X.k=z.p,report="Y")}
# function used to evaluate:
eval.fct <- function(yperm,yobs){return(as.numeric(yperm>=yobs))}
# calculate p-value:
p.perm <- (rowSums(apply(Y.perm,MARGIN=1,FUN=eval.fct,yobs=Yk.obs))+1)/(n.perm+1)
```

```
Landscape.fct <- function(X.k,report="ALL",inputtype="Z",logTransP="FALSE",alpha=0.05){
    # Leslie Foldager, Aarhus University, 6 March 2014
    # calculates Landscape for a sequence Z.k with various option for Z.k
    report <- toupper(report)
    if(report!="ALL" & report!="LESS" & report!="Y"){
        stop("report must be either ALL (default), LESS (Y and number of max segments) or Y")
    }
    inputtype <- toupper(inputtype)
    if(inputtype!="Z" & inputtype!="P" &
        inputtype!="LOGP" & inputtype!="MINUSLOGP"){
                stop("inputtype must be either Z (default), P, LOGP or MINUSLOGP")
    }
    logTransP <- toupper(logTransP)
    if(logTransP!="FALSE" & logTransP!="TRUE"){
        stop("logTransP must be either FALSE (default) or TRUE")
    }
    if(logTransP=="TRUE" & inputtype!="P"){
        stop("logTransP=TRUE assumes inputtype to be P")
    }
    if(inputtype=="Z"){
        Z.k <- X.k
    }
    if(logTransP=="TRUE" & inputtype=="P"){
        Z.k <- log(alpha)-log(X.k)
    }
    if(logTransP=="FALSE" & inputtype=="P"){
        Z.k <- ifelse(X.k<alpha,1,-1)
    }
    if(logTransP=="FALSE" & inputtype=="LOGP"){
        Z.k <- log(alpha)-X.k
    }
    if(logTransP=="FALSE" & inputtype=="MINUSLOGP"){
        Z.k <- log(alpha)+X.k
    }
    K <- length(Z.k)
    k.idx <- 1:K
    if(any(Z.k>0)){
        # --------------------------------
        # ---- Equation (2.1): U_{n,m} ----
        U.nm <- matrix(NA,nrow=K,ncol=K)
        for(m in k.idx){
            for(n in 1:m){
                U.nm[n,m] <- sum(Z.k[n:m])
            }
        }
        # --------------------------------
        # -----  Equation (2.4): A_k  -----
        A.k <- rep(0,K+2)
        for(k in (k.idx+1)){
            A.k[k] <- max(0,Z.k[k-1]+A.k[k-1])
        }
        A.k <- A.k[-1]
        # -------------------------------
        # --- Equation (2.5): sections ---
        ti0last <- t00 <- 0
        si0 <- ti0 <- list()
        i <- 0
        while(ti0last<(K-1)){
            i <- i+1
            idx <- which(A.k[(ti0last+1):(K+1)]>0)+ti0last
            if(suppressWarnings(any(idx))){
                si0[[i]] <- min(idx)
                idx <- which(A.k[si0[[i]]:(K+1)]==0)-1+(si0[[i]]-1)
                if(suppressWarnings(any(idx))){
                    ti0last <- ti0[[i]] <- min(idx)
                }
            }else{
                ti0last <- K
            }
    }
    si0 <- unlist(si0)
    ti0 <- unlist(ti0)
    lgt.I <- length(si0)
    sections <- cbind(si0,ti0)
    # --------------------------------------
    # --- Equation (2.6): init recursion ---
    ei0 <- Yi0 <- rep(NA,lgt.I)
    for(i in 1:lgt.I){
        Yi0[i] <- max(A.k[si0[i]:ti0[i]])
        ei0[i] <- min(which(A.k[si0[i]:ti0[i]]==Yi0[i])+(si0[i]-1))
    }
    # --------------------------------------
    # ----  recursion - equation (2.7)  ----
    dif <- c(NA,(A.k[-1])-A.k[-(K+1)])
    eij <- Yij <- tij <- sij <- list()
    for(i in 1:lgt.I){
        eijtmp <- Yijtmp <- sijtmp <- tijtmp <- list()
        j <- 1
        sijtmp[[j]] <- si0[i]
        tijtmp[[j]] <- ti0[i]
        Yijtmp[[j]] <- Yi0[i]
        eijtmp[[j]] <- ei0[i]
        if(eijtmp[[j]]<K){
            j <- j+1
            dif.idx <- which(dif[(eijtmp[[j-1]]+1):ti0[i]]>0)+eijtmp[[j-1]]
            if(suppressWarnings(any(dif.idx))){
                sijnext <- min(dif.idx)
                while(sijnext<=ti0[i]){
                    sijtmp[[j]] <- sijnext
                    dif.idx <- which((A.k[sijtmp[[j]]:ti0[i]]-A.k[sijtmp[[j]]-1])<=0)+(sijtmp[[j]]-1)
                    if(suppressWarnings(any(dif.idx))){
                        tijtmp[[j]] <-
                            min(which((A.k[sijtmp[[j]]:ti0[i]]-A.k[sijtmp[[j]]-1])<=0)+(sijtmp[[j]]-1))
                    }else{
                        tijtmp[[j]] <- ti0[i]
                    }
                    Yijtmp[[j]] <- max(A.k[sijtmp[[j]]:tijtmp[[j]]])
                    eijtmp[[j]] <-
                        min(which(A.k[sijtmp[[j]]:tijtmp[[j]]]==Yijtmp[[j]])+(sijtmp[[j]]-1))
                    j <- j+1
                    dif.idx <- which(dif[(eijtmp[[j-1]]+1):ti0[i]]>0)+eijtmp[[j-1]]
                    if(suppressWarnings(any(dif.idx))){
                        sijnext <- min(dif.idx)
                    }else{
                        sijnext <- K+1
                    }
                }
            }
        }
        sij[[i]] <- unlist(sijtmp)
        tij[[i]] <- unlist(tijtmp)
        Yij[[i]] <- unlist(Yijtmp)
        eij[[i]] <- unlist(eijtmp)
    }
    max.seg <- cbind(unlist(sij),unlist(eij))
    yk <- rep(0,length(Z.k))
    y <- unlist(Yij)
    ni <- nrow(max.seg)
    for(i in 1:ni){
        for(k in max.seg[i,1]:max.seg[i,2]){
            yk[k] <- y[i]
        }
    }
    N.max.segm <- nrow(max.seg)
    # --------------------------------------
}else{
    A.k <- rep(0,K+1)  # no Z.k>0
    N.max.segm <- 0
```

```
        yk <- rep(0,K)
        max.seg <- sections <- sij <- tij <- eij <- Yij <- NULL
    }
    if(report=="ALL"){
        return(result <- list(Z.k=Z.k,N.max.segm=N.max.segm,max.segm=max.seg,
            Y.k=yk,sect=sections,A.k=c(0,A.k),sij=sij,tij=tij,eij=eij,Yij=Yij))
    }else if(report=="LESS"){
        return(result <- list(N.max.segm=N.max.segm,Y.k=yk))
    }else{
        return(Y.k=yk)
    }
}
```

References

Ahdidan, J., Foldager, L., Rosenberg, R. et al. (2013). „Hippocampal volume and serotonin transporter polymorphism in major depressive disorder“. *Acta Neuropsychiatrica* **25**: 206–214.

Amato, R., Pinelli, M., D'Andrea, D. et al. (2010). „A novel approach to simulate gene-environment interactions in complex diseases“. *BMC Bioinformatics* **11**: 8.

Amemiya, T. (1984). „Tobit models - a survey“. *J. Econom.* **24**: 3–61.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders. DSM-IV*. Washington, DC: American Psychiatric Association,

Amos-Landgraf, J.M., Cottle, A., Plenge, R.M. et al. (2006). „X chromosome-inactivation patterns of 1,005 phenotypically unaffected females“. *Am. J. Hum. Genet.* **79**: 493–499.

Arendt, M., Rosenberg, R., Foldager, L. et al. (2005). „Cannabis-induced psychosis and subsequent schizophrenia-spectrum disorders: follow-up study of 535 incident cases“. *Br. J. Psychiatry* **187**: 510–515.

Armitage, P. (1955). „Tests for Linear Trends in Proportions and Frequencies“. *Biometrics* **11**: 375–386.

Arsenault-Lapierre, G., Kim, C. and Turecki, G. (2004). „Psychiatric diagnoses in 3275 suicides: a meta-analysis“. *BMC Psychiatry* **4**: 37.

Augui, S., Nora, E.P. and Heard, E. (2011). „Regulation of X-chromosome inactivation by the X-inactivation centre“. *Nat. Rev. Genet.* **12**: 429–442.

Balding, D.J. (2006). „A tutorial on statistical methods for population association studies“. *Nat. Rev. Genet.* **7**: 781–791.

Becker, T., Herold, C., Meesters, C. et al. (2011). „Significance Levels in Genome-Wide Interaction Analysis (GWIA)“. *Ann. Hum. Genet.* **75**: 29–35.

Benjamini, Y. and Hochberg, Y. (1995). „Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing“. *J R Stat Soc Series B Stat Methodol* **57**: 289–300.

Benros, M.E., Mortensen, P.B. and Eaton, W.W. (2012). „Autoimmune diseases and infections as risk factors for schizophrenia“. *Ann. N.Y. Acad. Sci.* **1262**: 56–66.

Benros, M.E., Nielsen, P.R., Nordentoft, M. et al. (2011). „Autoimmune diseases and severe infections as risk factors for schizophrenia: a 30-year population-based register study“. *Am. J. Psychiatry* **168**: 1303–1310.

Benros, M.E., Waltoft, B.L., Nordentoft, M. et al. (2013). „Autoimmune diseases and severe infections as risk factors for mood disorders: a nationwide study“. *JAMA Psychiatry* **70**: 812–820.

Benson, D.A., Clark, K., Karsch-Mizrachi, I. et al. (2014). „GenBank“. *Nucleic Acids Res.* **42**: D32–D37.

Bergen, S.E., O’Dushlaine, C.T., Ripke, S. et al. (2012). „Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder“. *Mol. Psychiatry* **17**: 880–886.

Boldt, A.B., Messias-Reason, I.J., Meyer, D. et al. (2010). „Phylogenetic nomenclature and evolution of mannose-binding lectin (MBL2) haplotypes“. *BMC Genet.* **11**: 38.

Borglum, A.D., Demontis, D., Grove, J. et al. (2013). „Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci“. *Mol. Psychiatry*, Epub ahead of print.

Borglum, A.D., Hampson, M., Kjeldsen, T.E. et al. (2001). „Dopa decarboxylase genotypes may influence age at onset of schizophrenia“. *Mol. Psychiatry* **6**: 712–717.

Breiman, L. (2001). „Random Forests“. *Machine Learning* **45**: 5–32.

Brent, D.A., Bridge, J., Johnson, B.A. and Connolly, J. (1996). „Suicidal behavior runs in families. A controlled family study of adolescent suicide victims“. *Arch. Gen. Psychiatry* **53**: 1145–1152.

Brown, A.S., Schaefer, C.A., Quesenberry C.P., Jr. et al. (2005). „Maternal exposure to toxoplasmosis and risk of schizophrenia in adult offspring“. *Am. J. Psychiatry* **162**: 767–773.

Browning, S.R. and Browning, B.L. (2007). „Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering“. *Am. J. Hum. Genet.* **81**: 1084–1097.

Buka, S.L., Tsuang, M.T., Torrey, E.F. et al. (2001). „Maternal infections and subsequent psychosis among offspring“. *Arch. Gen. Psychiatry* **58**: 1032–1037.

Buttenschøn, H.N., Flint, T.J., Foldager, L. et al. (2013). „An association study of suicide and candidate genes in the serotonergic system“. *J. Affect. Disord.* **148**: 291–298.

Buttenschøn, H.N., Foldager, L., Flint, T.J. et al. (2010). „Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3“. *Psychiatr. Genet.* **20**: 93–101.

Calle, M.L., Urrea, V., Malats, N. and Van Steen, K. (2008). *MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data*. Tech. rep. Universitat de Vic, 1–14.

Calle, M.L., Urrea, V., Malats, N. and Van Steen, K. (2010). „mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits“. *Bioinformatics* **26**: 2198–2199.

Carrel, L. and Willard, H.F. (2005). „X-inactivation profile reveals extensive variability in X-linked gene expression in females“. *Nature* **434**: 400–404.

Caspi, A., Hariri, A.R., Holmes, A. et al. (2010). „Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits“. *Am. J. Psychiatry* **167**: 509–527.

Castagnini, A. and Foldager, L. (2013). „Variations in incidence and age of onset of acute and transient psychotic disorders“. *Soc. Psychiatry Psychiatr. Epidemiol.* **48**: 1917–1922.

Cattaert, T., Calle, M.L., Dudek, S.M. et al. (2011). „Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise“. *Ann. Hum. Genet.* **75**: 78–89.

Cerny, V. (1985). „Thermodynamical Approach to the Traveling Salesman Problem - An Efficient Simulation Algorithm“. *J. Optim. Theory Appl.* **45**: 41–51.

Chen, J., Song, Y., Yang, J. et al. (2013). „The contribution of TNF-alpha in the amygdala to anxiety in mice with persistent inflammatory pain“. *Neurosci. Lett.* **541**: 275–280.

Clayton, D. (2008). „Testing for association on the X chromosome“. *Biostatistics* **9**: 593–600.

Clayton, D.G. (2009). „Sex chromosomes and genetic association studies“. *Genome Med.* **1**: 110.

Cochran, W.G. (1954). „Some methods for strengthening the common chi-squared tests“. *Biometrics* **10**: 417–451.

Conover, W.J. (1999). *Practical nonparametric statistics*. New York: Wiley,

Courtet, P., Jollant, F., Buresi, C. et al. (2005). „The monoamine oxidase A gene may influence the means used in suicide attempts“. *Psychiatr. Genet.* **15**: 189–193.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013a). „Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs“. *Nat. Genet.* **45**: 984–994.

Cross-Disorder Group of the Psychiatric Genomics Consortium (2013b). „Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis“. *Lancet* **381**: 1371–1379.

Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row,

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press,

Deckert, J., Catalano, M., Syagailo, Y.V. et al. (1999). „Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder“. *Hum. Mol. Genet.* **8**: 621–624.

Degn, B., Lundorf, M.D., Wang, A. et al. (2001). „Further evidence for a bipolar risk gene on chromosome 12q24 suggested by investigation of haplotype sharing and allelic association in patients from the Faroe Islands“. *Mol. Psychiatry* **6**: 450–455.

Dommett, R.M., Klein, N. and Turner, M.W. (2006). „Mannose-binding lectin in innate immunity: past, present and future“. *Tissue Antigens* **68**: 193–209.

Donohoe, G., Walters, J., Hargreaves, A. et al. (2013). „Neuropsychological effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253“. *Genes Brain Behav.* **12**: 203–209.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003). „Multiple hypothesis testing in microarray experiments“. *Statist. Sci.* **18**: 71–103.

Eaton, W.W., Byrne, M., Ewald, H. et al. (2006). „Association of schizophrenia and autoimmune diseases: linkage of Danish national registers“. *Am. J. Psychiatry* **163**: 521–528.

Eaton, W.W., Pedersen, M.G., Nielsen, P.R. and Mortensen, P.B. (2010). „Autoimmune diseases, bipolar disorder, and non-affective psychosis“. *Bipolar Disord.* **12**: 638–646.

Elfving, B., Buttenschon, H.N., Foldager, L. et al. (2012). „Depression, the Val66Met polymorphism, age, and gender influence the serum BDNF level“. *J. Psychiatr. Res.* **46**: 1118–1125.

Elfving B. and Buttenschøn, H.N., Foldager, L., Poulsen, P.H.P. et al. (2014). „Depression and BMI influences the serum Vascular Endothelial Growth Factor level". *Int. J. Neuropsychopharmacol.* Epub ahead of print.

ENCODE Project Consortium, Bernstein, B.E., Birney, E. et al. (2012). „An integrated encyclopedia of DNA elements in the human genome". *Nature* **489**: 57–74.

Endicott, J. and Spitzer, R.L. (1978). „A diagnostic interview: the schedule for affective disorders and schizophrenia". *Arch. Gen. Psychiatry* **35**: 837–844.

Erhardt, A., Akula, N., Schumacher, J. et al. (2012). „Replication and meta-analysis of TMEM132D gene variants in panic disorder". *Transl. Psychiatry* **2**: e156.

Erhardt, A., Czibere, L., Roeske, D. et al. (2011). „TMEM132D, a new candidate for anxiety phenotypes: evidence from human and mouse studies". *Mol. Psychiatry* **16**: 647–663.

Ewald, H., Degn, B., Mors, O. and Kruse, T.A. (1998). „Significant linkage between bipolar affective disorder and chromosome 12q24". *Psychiatr. Genet.* **8**: 131–140.

Ewald, H., Flint, T., Kruse, T.A. and Mors, O. (2002). „A genome-wide scan shows significant linkage between bipolar disorder and chromosome 12q24.3 and suggestive linkage to chromosomes 1p22-21, 4p16, 6q14-22, 10q26 and 16p13.3". *Mol. Psychiatry* **7**: 734–744.

Faresjö, T. and Faresjö, A. (2010). „To match or not to match in epidemiological studies–same outcome but less power". *Int. J. Environ. Res. Public Health* **7**: 325–332.

Ferrari, A.J., Baxter, A.J. and Whiteford, H.A. (2011). „A systematic review of the global distribution and availability of prevalence data for bipolar disorder". *J. Affect. Disord.* **134**: 1–13.

Ferrari, A.J., Saha, S., McGrath, J.J. et al. (2012). „Health states for schizophrenia and bipolar disorder within the Global Burden of Disease 2010 Study". *Popul. Health Metr.* **10**: 16.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A. et al. (2008). „Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder". *Nat. Genet.* **40**: 1056–1058.

Fillman, S.G., Cloonan, N., Catts, V.S. et al. (2013). „Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia". *Mol. Psychiatry* **18**: 206–214.

Fisher, RA. (1932). *Statistical methods for research workers*. Oliver and Boyd.

Foldager, L., Als, T.D. and Grove, J. (2013). „Comparison of methods for genome-wide gene-environment interaction analysis". In: *Abstract book for XXI World Congress of Psychiatric Genetics (WCPG)*. International Society of Psychiatric Genetics (ISPG). Boston, MA, USA, 279–280.

Foldager, L., Buttenschøn, H.N., Flint, T.J. et al. (2009a). „Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3". In: *Abstract book for XVII World Congress on Psychiatric Genetics (WCPG)*. International Society of Psychiatric Genetics (ISPG). San Diego, CA, USA, 53.

Foldager, L., Köhler, O., Steffensen, R. et al. (2014). „Bipolar and panic disorders may be associated with hereditary defects in the innate immune system". *J. Affect. Disord.* **164**: 148–154.

Foldager, L., Pedersen, C.B., Nyegaard, M. et al. (2010). „Conditional logic regression: identifying snp interactions from individually time-matched case-control data."

In: *Abstract book for XVIII World Congress on Psychiatric Genetics (WCPG)*. International Society on Psychiatric Genetics (ISPG). Athens, Greece, 202–203.

Foldager, L., Steffensen, R., Thiel, S. et al. (2009b). „Are defects in the innate immune defense contributing factors for mental disorders?" In: *Abstract book for XVI World Congress on Psychiatric Genetics (WCPG)*. International Society of Psychiatric Genetics (ISPG). Osaka, Japan, 225.

Foldager, L., Steffensen, R., Thiel, S. et al. (2012). „MBL and MASP-2 concentrations in serum and MBL2 promoter polymorphisms are associated to schizophrenia". *Acta Neuropsychiatrica* **24**: 199–207.

Forman, J., Taruscio, D., Llera, V.A. et al. (2012). „The need for worldwide policy and action plans for rare diseases". *Acta Paediatr.* **101**: 805–807.

Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009). „Human genetic variation and its contribution to complex traits". *Nat. Rev. Genet.* **10**: 241–251.

Fromer, M., Pocklington, A.J., Kavanagh, D.H. et al. (2014). „De novo mutations in schizophrenia implicate synaptic networks". *Nature* **506**: 179–184.

Galli, L., Chiappini, E. and de, Martino M. (2012). „Infections and autoimmunity". *Pediatr. Infect. Dis. J.* **31**: 1295–1297.

Garred, P., Honore, C., Ma, Y.J. et al. (2009). „MBL2, FCN1, FCN2 and FCN3-The genes behind the initiation of the lectin pathway of complement". *Mol. Immunol.* **46**: 2737–2744.

Garred, P., Larsen, F., Seyfarth, J. et al. (2006). „Mannose-binding lectin and its genetic variants". *Genes Immun.* **7**: 85–94.

Gaunt, T.R., Rodriguez, S., Zapata, C. and Day, I.N. (2006). „MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers". *BMC Bioinformatics* **7**: 227.

Gonda, X., Fountoulakis, K.N., Harro, J. et al. (2011). „The possible contributory role of the S allele of 5-HTTLPR in the emergence of suicidality". *J. Psychopharmacol.* **25**: 857–866.

Gonda, X., Pompili, M., Serafini, G. et al. (2012). „Suicidal behavior in bipolar disorder: epidemiology, characteristics and major risk factors". *J. Affect. Disord.* **143**: 16–26.

Guedj, M., Nuel, G. and Prum, B. (2008). „A note on allelic tests in case-control association studies". *Ann. Hum. Genet.* **72**: 407–409.

Guedj, M., Wojcik, J., Chiesa, E. la et al. (2006). „A fast, unbiased and exact allelic test for case-control association studies". *Hum. Hered.* **61**: 210–221.

Hashimoto, R., Ikeda, M., Ohi, K. et al. (2013). „Genome-wide association study of cognitive decline in schizophrenia". *Am. J. Psychiatry* **170**: 683–684.

Hastings, W.K. (1970). „Monte-Carlo Sampling Methods Using Markov Chains and Their Applications". *Biometrika* **57**: 97–109.

Havik, B., Le, Hellard S., Rietschel, M. et al. (2011). „The complement control-related genes CSMD1 and CSMD2 associate to schizophrenia". *Biol. Psychiatry* **70**: 35–42.

Heitzeneder, S., Seidel, M., Forster-Waldl, E. and Heitger, A. (2012). „Mannan-binding lectin deficiency - good news, bad news, doesn't matter?" *Clin. Immunol.* **143**: 22–38.

Hoban, S., Bertorelle, G. and Gaggiotti, O.E. (2012). „Computer simulations: tools for population and evolutionary genetics". *Nat. Rev. Genet.* **13**: 110–122.

Hollegaard, M.V., Grove, J., Grauholm, J. et al. (2011). „Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source“. *BMC Genet.* **12**: 58.

Holmans, P., Moskvina, V., Jones, L. et al. (2013). „A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson’s disease“. *Hum. Mol. Genet.* **22**: 1039–1049.

Hommel, G. (1988). „A stagewise rejective multiple test procedure based on a modified Bonferroni test“. *Biometrika* **75**: 383–386.

Hope, A.C.A. (1968). „A simplified Monte Carlo significance test procedure“. *Journal of the Royal Statistical Society, Series B, Methodological* **30**: 582–598.

Howie, B., Marchini, J. and Stephens, M. (2011). „Genotype imputation with thousands of genomes“. *G3.(Bethesda.)* **1**: 457–470.

Howie, B.N., Donnelly, P. and Marchini, J. (2009). „A flexible and accurate genotype imputation method for the next generation of genome-wide association studies“. *PLoS Genet.* **5**: e1000529.

Hu, X.Z., Lipsky, R.H., Zhu, G. et al. (2006). „Serotonin transporter promoter gain-of-function genotypes are linked to obsessive-compulsive disorder“. *Am. J. Hum. Genet.* **78**: 815–826.

International HapMap 3 Consortium (2010). „Integrating common and rare genetic variation in diverse human populations“. *Nature* **467**: 52–58.

International HapMap Consortium (2005). „A haplotype map of the human genome“. *Nature* **437**: 1299–1320.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G. et al. (2007). „A second generation human haplotype map of over 3.1 million SNPs“. *Nature* **449**: 851–861.

International Schizophrenia Consortium, Wray, N.R., Stone, J.L. et al. (2009). „Common polygenic variation contributes to risk of schizophrenia and bipolar disorder“. *Nature* **460**: 748–752.

Izbicki, R., Fossaluza, V., Hounie, A.G. et al. (2012). „Testing allele homogeneity: the problem of nested hypotheses“. *BMC Genet.* **13**: 103.

Joo, J., Kwak, M., Ahn, K. and Zheng, G. (2009). „A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium“. *Biometrics* **65**: 1115–1122.

Kalsi, G., McQuillin, A., Degn, B. et al. (2006). „Identification of the Slynar gene (AY070435) and related brain expressed sequences as a candidate gene for susceptibility to affective disorders through allelic and haplotypic association with bipolar disorder on chromosome 12q24“. *Am. J. Psychiatry* **163**: 1767–1776.

Karlin, S. and Altschul, S.F. (1990). „Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.“ *Proc Natl Acad Sci U S A* **87**: 2264–2268.

Karlin, S. and Dembo, A. (1992). „Limit Distributions of Maximal Segmental Score among Markov-Dependent Partial Sums“. *Advances in Applied Probability* **24**: 113–140.

Kent, W.J., Sugnet, C.W., Furey, T.S. et al. (2002). „The human genome browser at UCSC“. *Genome Res.* **12**: 996–1006.

Kirkpatrick, B. and Miller, B.J. (2013). „Inflammation and schizophrenia“. *Schizophr. Bull.* **39**: 1174–1179.

Kirkpatrick, S., Gelatt C.D., Jr. and Vecchi, M.P. (1983). „Optimization by simulated annealing“. *Science* **220**: 671–680.

Knapp, M. (2001). „Re: ”Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions”“. *Am. J. Epidemiol.* **154**: 287–288.

Knapp, M. (2008). „On the asymptotic equivalence of allelic and trend statistic under Hardy-Weinberg equilibrium“. *Ann. Hum. Genet.* **72**: 589.

Kolstad, H.A., Hansen, A.M., Kaergaard, A. et al. (2011). „Job strain and the risk of depression: is reporting biased?“ *Am. J. Epidemiol.* **173**: 94–102.

Kooperberg, C. and Ruczinski, I. (2005). „Identifying interacting SNPs using Monte Carlo logic regression“. *Genet. Epidemiol.* **28**: 157–170.

Kooperberg, C., Ruczinski, I., LeBlanc, M.L. and Hsu, L. (2001). „Sequence analysis using logic regression“. *Genet. Epidemiol.* **21 Suppl 1**: S626–S631.

Kwon, E., Wang, W. and Tsai, L.H. (2013). „Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets“. *Mol. Psychiatry* **18**: 11–12.

Lambert, J.C., Grenier-Boley, B., Chouraki, V. et al. (2010). „Implication of the immune system in Alzheimer’s disease: evidence from genome-wide pathway analysis“. *J. Alzheimers. Dis.* **20**: 1107–1118.

Larsen, J.K., Bendsen, B.B, Foldager, L. and Munk-Jorgensen, P. (2010). „Prematurity and low birth weight as risk factors for the development of affective disorder, especially depression and schizophrenia: a register study.“ *Acta Neuropsychiatrica* **22**: 284–291.

Leboyer, M., Soreca, I., Scott, J. et al. (2012). „Can bipolar disorder be viewed as a multi-system inflammatory disease?“ *J. Affect. Disord.* **141**: 1–10.

Lee, S.H., DeCandia, T.R., Ripke, S. et al. (2012). „Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs“. *Nat. Genet.* **44**: 247–250.

Lee, S.H., Wray, N.R., Goddard, M.E. and Visscher, P.M. (2011). „Estimating missing heritability for disease from genome-wide association studies“. *Am. J. Hum. Genet.* **88**: 294–305.

Li, D. and He, L. (2006). „Further clarification of the contribution of the tryptophan hydroxylase (TPH) gene to suicidal behavior using systematic allelic and genotypic meta-analyses“. *Hum. Genet.* **119**: 233–240.

Li, H. and Homer, N. (2010a). „A survey of sequence alignment algorithms for next-generation sequencing“. *Brief. Bioinform.* **11**: 473–483.

Li, Q., Louis, T.A., Fallin, M.D. and Ruczinski, I. (2009). *Trio Logic Regression - Detection of SNP-SNP Interactions in Case-Parent Trios*. Tech. rep., 1–43.

Li, Y., Willer, C.J., Ding, J. et al. (2010b). „MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes“. *Genet. Epidemiol.* **34**: 816–834.

Llinares, J. (2010). „A regulatory overview about rare diseases“. *Adv. Exp. Med. Biol.* **686**: 193–207.

Loley, C., Konig, I.R., Hothorn, L. and Ziegler, A. (2013). „A unifying framework for robust association testing, estimation, and genetic model selection using the generalized linear model“. *Eur. J. Hum. Genet.* **21**: 1442–1448.

Louis, T.A. and Zeger, S.L. (2009). „Effective communication of standard errors and confidence intervals“. *Biostatistics* **10**: 1–2.

Lublin, H., Eberhard, J. and Levander, S. (2005). „Current therapy issues and unmet clinical needs in the treatment of schizophrenia: a review of the new generation antipsychotics“. *Int. Clin. Psychopharmacol.* **20**: 183–198.

Mahachie John, J.M., Cattaert, T., Van Lishout, F. et al. (2012). „Lower-order effects adjustment in quantitative traits model-based multifactor dimensionality reduction“. *PLoS One* **7**: e29594.

Mahachie John, J.M., Van Lishout, F. and Van Steen, K. (2011). „Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data“. *Eur. J. Hum. Genet.* **19**: 696–703.

Mangs, A.H. and Morris, B.J. (2007). „The Human Pseudoautosomal Region (PAR): Origin, Function and Future“. *Curr. Genomics* **8**: 129–136.

Marchini, J. and Howie, B. (2010). „Genotype imputation for genome-wide association studies“. *Nat. Rev. Genet.* **11**: 499–511.

Matsushita, M. (2010). „Ficolins: complement-activating lectins involved in innate immunity“. *J. Innate. Immun.* **2**: 24–32.

Mayilyan, K.R. (2012). „Complement genetics, deficiencies, and disease associations“. *Protein Cell* **3**: 487–496.

Mayilyan, K.R., Weinberger, D.R. and Sim, R.B. (2008). „The complement system in schizophrenia“. *Drug News Perspect.* **21**: 200–210.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R. et al. (2008). „Genome-wide association studies for complex traits: consensus, uncertainty and challenges“. *Nat. Rev. Genet.* **9**: 356–369.

McClellan, J.M., Susser, E. and King, M.C. (2007). „Schizophrenia: a common disease caused by multiple rare alleles“. *Br. J. Psychiatry* **190**: 194–199.

McGrath, J., Saha, S., Chant, D. and Welham, J. (2008). „Schizophrenia: a concise overview of incidence, prevalence, and mortality“. *Epidemiol. Rev.* **30**: 67–76.

McGrath, J.J., Petersen, L., Agerbo, E. et al. (2014). „A comprehensive assessment of parental age and psychiatric disorders“. *JAMA Psychiatry* **71**: 301–309.

McGuffin, P., Owen, M.J. and Farmer, A.E. (1995). „Genetic basis of schizophrenia“. *Lancet* **346**: 678–682.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N. et al. (1953). „Equation of State Calculations by Fast Computing Machines“. *J Chem Phys* **21**: 1087–1092.

Meyfroidt, G., Guiza, F., Ramon, J. and Bruynooghe, M. (2009). „Machine learning techniques to examine large patient databases“. *Best. Pract. Res. Clin. Anaesthesiol.* **23**: 127–143.

Mitchell, K.J. and Porteous, D.J. (2011). „Rethinking the genetic architecture of schizophrenia“. *Psychol. Med.* **41**: 19–32.

Mors, O., Perto, G.P. and Mortensen, P.B. (2011). „The Danish Psychiatric Central Research Register“. *Scand. J. Public Health* **39**: 54–57.

Mortensen, P.B., Norgaard-Pedersen, B., Waltoft, B.L. et al. (2007). „Early Infections of Toxoplasma gondii and the Later Development of Schizophrenia“. *Schizophr. Bull.* **33**: 741–744.

Mortensen, P.B., Pedersen, C.B., Hougaard, D.M. et al. (2010). „A Danish National Birth Cohort study of maternal HSV-2 antibodies as a risk factor for schizophrenia in their offspring“. *Schizophr. Res.* **122**: 257–263.

Motsinger, A.A. and Ritchie, M.D. (2006). „Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies“. *Hum. Genomics* **2**: 318–328.

Murcray, C.E., Lewinger, J.P. and Gauderman, W.J. (2009). „Gene-environment interaction in genome-wide association studies“. *Am. J. Epidemiol.* **169**: 219–226.

Murray, C.J., Ezzati, M., Flaxman, A.D. et al. (2012). „GBD 2010: design, definitions, and metrics“. *Lancet* **380**: 2063–2066.

NCBI Resource Coordinators (2014). „Database resources of the National Center for Biotechnology Information“. *Nucleic Acids Res.* **42**: D7–D17.

Ng, M.Y., Levinson, D.F., Faraone, S.V. et al. (2009). „Meta-analysis of 32 genome-wide linkage studies of schizophrenia“. *Mol. Psychiatry* **14**: 774–785.

Nicodemus, K.K., Marenco, S., Batten, A.J. et al. (2008). „Serious obstetric complications interact with hypoxia-regulated/vascular-expression genes to influence schizophrenia risk“. *Mol. Psychiatry* **13**: 873–877.

Nielsen, P.R., Benros, M.E. and Mortensen, P.B. (2013a). „Hospital Contacts With Infection and Risk of Schizophrenia: A Population-Based Cohort Study With Linkage of Danish National Registers“. *Schizophr. Bull.* Epub ahead of print.

Nielsen, P.R., Laursen, T.M. and Mortensen, P.B. (2013b). „Association between parental hospital-treated infection and the risk of schizophrenia in adolescence and early adulthood“. *Schizophr. Bull.* **39**: 230–237.

Nielsen, P.R., Mortensen, P.B., Dalman, C. et al. (2013c). „Fetal growth and schizophrenia: a nested case-control and case-sibling study“. *Schizophr. Bull.* **39**: 1337–1342.

Norgaard-Pedersen, B. and Hougaard, D.M. (2007). „Storage policies and use of the Danish Newborn Screening Biobank“. *J.Inherit.Metab Dis.* **30**: 530–536.

Nyegaard, M., Demontis, D., Foldager, L. et al. (2010). „CACNA1C (rs1006737) is associated with schizophrenia“. *Mol. Psychiatry* **15**: 119–121.

Nyholt, D.R. (2004). „A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other“. *Am. J. Hum. Genet.* **74**: 765–769.

Ohno, S. (1967). *Sex chromosomes and sex linked genes*. Berlin, Germany: Springer Verlag,

Olesen, H.V., Jensenius, J.C., Steffensen, R. et al. (2006). „The mannan-binding lectin pathway and lung disease in cystic fibrosis - dysfunction of mannan-binding lectin-associated serine protease 2 (MASP-2) may be a major modifier“. *Clin. Immunol.* **121**: 324–331.

Otowa, T., Yoshida, E., Sugaya, N. et al. (2009). „Genome-wide association study of panic disorder in the Japanese population“. *J.Hum.Genet.* **54**: 122–126.

Pandey, G.N. (2013). „Biological basis of suicide and suicidal behavior“. *Bipolar Disord.* **15**: 524–541.

Parsey, R.V., Hastings, R.S., Oquendo, M.A. et al. (2006). „Effect of a triallelic functional polymorphism of the serotonin-transporter-linked promoter region on expression of serotonin transporter in the human brain“. *Am. J. Psychiatry* **163**: 48–51.

Pedersen, C.B., Gotzsche, H., Moller, J.O. and Mortensen, P.B. (2006). „The Danish Civil Registration System. A cohort of eight million persons“. *Dan. Med. Bull.* **53**: 441–449.

Peng, B. and Amos, C.I. (2010). „Forward-time simulation of realistic samples for genome-wide association studies“. *BMC Bioinformatics* **11**: 442.

Peng, B., Chen, H.S., Mechanic, L.E. et al. (2013). „Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators“. *Bioinformatics* **29**: 1101–1102.

Peng, B. and Kimmel, M. (2005). „simuPOP: a forward-time population genetics simulation environment“. *Bioinformatics* **21**: 3686–3687.

Peng, B., Kimmel, M. and Amos, C.I. (2012). *Forward-Time Population Genetics Simulations: Methods, Implementation, and Applications*. New Jersey, USA: Wiley-Blackwell,

Pereira, A.C., McQuillin, A., Puri, V. et al. (2011). „Genetic association and sequencing of the insulin-like growth factor 1 gene in bipolar affective disorder“. *Am. J. Med. Genet. B NeuroPsychiatr. Genet.* **156**: 177–187.

Petersen, L., Mortensen, P.B. and Pedersen, C.B. (2011). „Paternal age at birth of first child and risk of schizophrenia“. *Am. J. Psychiatry* **168**: 82–88.

Petersen, L., Sorensen, T.I., Andersen, P.K. et al. (2013). „Genetic and familial environmental effects on suicide–an adoption study of siblings“. *PLoS One* **8**: e77973.

Pinelli, M., Scala, G., Amato, R. et al. (2012). „Simulating gene-gene and gene-environment interactions in complex diseases: Gene-Environment iNteraction Simulator 2“. *BMC Bioinformatics* **13**: 132.

Pinsonneault, J.K., Papp, A.C. and Sadee, W. (2006). „Allelic mRNA expression of X-linked monoamine oxidase a (MAOA) in human brain: dissection of epigenetic and genetic factors“. *Hum. Mol. Genet.* **15**: 2636–2649.

Pruitt, K.D., Brown, G.R., Hiatt, S.M. et al. (2014). „RefSeq: an update on mammalian reference sequences“. *Nucleic Acids Res.* **42**: D756–D763.

Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011). „Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4“. *Nat. Genet.* **43**: 977–983.

Purcell, S.M., Moran, J.L., Fromer, M. et al. (2014). „A polygenic burden of rare disruptive mutations in schizophrenia“. *Nature* **506**: 185–190.

Qin, P. (2011). „The impact of psychiatric illness on suicide: differences by diagnosis of disorders and by sex and age of subjects“. *J. Psychiatr. Res.* **45**: 1445–1452.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Ramanan, V.K. and Saykin, A.J. (2013). „Pathways to neurodegeneration: mechanistic insights from GWAS in Alzheimer’s disease, Parkinson’s disease, and related disorders“. *Am. J. Neurodegener. Dis.* **2**: 145–175.

Riley, B. and Kendler, K.S. (2006). „Molecular genetic studies of schizophrenia“. *Eur. J. Hum. Genet.* **14**: 669–680.

Ripke, S., O’Dushlaine, C., Chambert, K. et al. (2013). „Genome-wide association analysis identifies 13 new risk loci for schizophrenia“. *Nat. Genet.* **45**: 1150–1159.

Ripke, S., Sanders, A.R., Kendler, K.S. et al. (2011). „Genome-wide association study identifies five new schizophrenia loci“. *Nat. Genet.* **43**: 969–976.

Ritchie, M.D., Hahn, L.W., Roodi, N. et al. (2001). „Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer“. *Am. J. Hum. Genet.* **69**: 138–147.

Ruczinski, I. (2000). „Logic regression and statistical issues related to the protein folding problem“. PhD thesis. Seattle: University of Washington, Dept. of Statistics.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003). „Logic regression“. *J. Comput. Graph. Stat.* **12**: 475–511.

Sørensen, R., Thiel, S. and Jensenius, J.C. (2005). „Mannan-binding-lectin-associated serine proteases, characteristics and disease associations“. *Springer Semin. Immunopathol.* **27**: 299–319.

Saha, S., Chant, D. and McGrath, J. (2007). „A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time?“ *Arch. Gen. Psychiatry* **64**: 1123–1131.

Saha, S., Chant, D., Welham, J. and McGrath, J. (2005). „A systematic review of the prevalence of schizophrenia“. *PLoS Med.* **2**: e141.

Salazar, A., Gonzalez-Rivera, B.L., Redus, L. et al. (2012). „Indoleamine 2,3-dioxygenase mediates anhedonia and anxiety-like behaviors caused by peripheral lipopolysaccharide immune challenge“. *Horm. Behav.* **62**: 202–209.

Samuel, A.L. (1959). „Some Studies in Machine Learning Using the Game of Checkers“. *IBM Journal* **3**: 210–229.

Sarma, J.V. and Ward, P.A. (2011). „The complement system“. *Cell Tissue Res.* **343**: 227–235.

Sasieni, P.D. (1997). „From genotypes to genes: doubling the sample size“. *Biometrics* **53**: 1253–1261.

Schaid, D.J. and Jacobsen, S.J. (1999). „Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions“. *Am. J. Epidemiol.* **149**: 706–711.

Schaid, D.J., Rowland, C.M., Tines, D.E. et al. (2002). „Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous“. *Am. J. Hum. Genet.* **70**: 425–434.

Schumacher, J., Kristensen, A.S., Wendland, J.R. et al. (2011). „The genetics of panic disorder“. *J. Med. Genet.* **48**: 361–368.

Schwender, H. (2007). *Minimization of Boolean Expressions Using Matrix Algebra*. Tech. rep., 1–29.

Schwender, H., Bowers, K., Fallin, M.D. and Ruczinski, I. (2011a). „Importance measures for epistatic interactions in case-parent trios“. *Ann. Hum. Genet.* **75**: 122–132.

Schwender, H. and Ickstadt, K. (2008). „Identification of SNP interactions using logic regression“. *Biostatistics* **9**: 187–198.

Schwender, H. and Ruczinski, I. (2010). „Logic regression and its extensions“. *Adv. Genet.* **72**: 25–45.

Schwender, H., Ruczinski, I. and Ickstadt, K. (2011b). „Testing SNPs and sets of SNPs for importance in association studies“. *Biostatistics* **12**: 18–32.

Severinsen, J.E., Bjarkam, C.R., Kiar-Larsen, S. et al. (2006). „Evidence implicating BRD1 with brain development and susceptibility to both schizophrenia and bipolar affective disorder“. *Mol. Psychiatry* **11**: 1126–1138.

Shaffer, J.P. (1995). „Multiple hypothesis testing“. *Annu. Rev. Psychol.* **46**: 561–584.

Sherry, S.T., Ward, M.H., Kholodov, M. et al. (2001). „dbSNP: the NCBI database of genetic variation“. *Nucleic Acids Res.* **29**: 308–311.

Shi, J., Levinson, D.F., Duan, J. et al. (2009). „Common variants on chromosome 6p22.1 are associated with schizophrenia“. *Nature* **460**: 753–757.

Simon, N.M., Otto, M.W., Wisniewski, S.R. et al. (2004). „Anxiety disorder comorbidity in bipolar disorder patients: data from the first 500 participants in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD)“. *Am. J. Psychiatry* **161**: 2222–2229.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J. et al. (2003). „The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes“. *Nature* **423**: 825–837.

Sklar, P., Smoller, J.W., Fan, J. et al. (2008). „Whole-genome association study of bipolar disorder“. *Mol. Psychiatry* **13**: 558–569.

Sorensen, H.J., Foldager, L., Roge, R. et al. (2013). „An association between autumn birth and clozapine treatment in patients with schizophrenia: A population-based analysis“. *Nord. J. Psychiatry*, Epub ahead of print.

Stabellini, R., Vasques, L.R., Mello, J.C. de et al. (2009). „MAOA and GYG2 are submitted to X chromosome inactivation in human fibroblasts“. *Epigenetics.* **4**: 388–393.

Stefansson, H., Ophoff, R.A., Steinberg, S. et al. (2009). „Common variants conferring risk of schizophrenia“. *Nature* **460**: 744–747.

Su, Z., Marchini, J. and Donnelly, P. (2011). „HAPGEN2: simulation of multiple disease SNPs“. *Bioinformatics* **27**: 2304–2305.

Sullivan, P.F., Daly, M.J. and O’Donovan, M. (2012). „Genetic architectures of psychiatric disorders: the emerging picture and its implications“. *Nat. Rev. Genet.* **13**: 537–551.

Tambuyzer, E. (2010). „Rare diseases, orphan drugs and their regulation: questions and misconceptions“. *Nat. Rev. Drug Discov.* **9**: 921–929.

Tamminga, C.A. and Holcomb, H.H. (2005). „Phenotype of schizophrenia: a review and formulation“. *Mol. Psychiatry* **10**: 27–39.

The 1000 Genomes Consortium, Abecasis, G.R., Altshuler, D.L. et al. (2010). „A map of human genome variation from population-scale sequencing“. *Nature* **467**: 1061–1073.

Thiel, S., Frederiksen, P.D. and Jensenius, J.C. (2006). „Clinical manifestations of mannan-binding lectin deficiency“. *Mol. Immunol.* **43**: 86–96.

Torrey, E.F., Bartko, J.J. and Yolken, R.H. (2012). „Toxoplasma gondii and other risk factors for schizophrenia: an update“. *Schizophr. Bull.* **38**: 642–647.

Troeng, T., Bergqvist, D. and Janson, L. (1994). „Incidence and causes of adverse outcomes of operation for chronic ischaemia of the leg“. *Eur. J. Surg.* **160**: 17–25.

Tsai, S.J., Hong, C.J. and Liou, Y.J. (2011). „Recent molecular genetic studies and methodological issues in suicide research“. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **35**: 809–817.

Turecki, G. (2001). „Suicidal behavior: is there a genetic predisposition?“ *Bipolar Disord.* **3**: 335–349.

Van Lishout, F., Mahachie John, J.M., Gusareva, E.S. et al. (2013). „An efficient algorithm to perform multiple testing in epistasis screening“. *BMC Bioinformatics* **14**: 138.

Veerappa, A.M., Padakannaya, P. and Ramachandra, N.B. (2013). „Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome“. *Funct. Integr. Genomics* **13**: 285–293.

Visscher, P.M., Brown, M.A., McCarthy, M.I. and Yang, J. (2012). „Five years of GWAS discovery“. *Am. J. Hum. Genet.* **90**: 7–24.

Voorman, A., Rice, K. and Lumley, T. (2012). „Fast computation for genome-wide association studies using boosted one-step statistics“. *Bioinformatics* **28**: 1818–1822.

Voracek, M. and Loibl, L.M. (2007). „Genetics of suicide: a systematic review of twin studies“. *Wien. Klin. Wochenschr.* **119**: 463–475.

Wan, X., Yang, C., Yang, Q. et al. (2010). „BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies“. *Am. J. Hum. Genet.* **87**: 325–340.

Wang, K. (2012). „Statistical tests of genetic association for case-control study designs“. *Biostatistics* **13**: 724–733.

Wellcome Trust Case Control Consortium (2007). „Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls“. *Nature* **447**: 661–678.

Westfall, P.H. and Young, S.S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons,

Willour, V.L., Seifuddin, F., Mahon, P.B. et al. (2012). „A genome-wide association study of attempted suicide“. *Mol. Psychiatry* **17**: 433–444.

Wing, J.K., Sartorius, N. and Ustun, T.B. (1998). *Diagnosis and Clinical Measurement in Psychiatry. A Reference Manual for SCAN*. Cambridge: Cambridge University Press,

Winham, S.J. and Motsinger-Reif, A.A. (2011). „An R package implementation of multifactor dimensionality reduction“. *BioData Min.* **4**: 24.

World Health Organization (1993). *The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research*. Geneva: World Health Organization,

Wright, S. (1938). „Size of population and breeding structure in relation to evolution“. *Science* **87**: 430–431.

Xiao, J.C., Buka, S.L., Cannon, T.D. et al. (2009). „Serological pattern consistent with infection with type I Toxoplasma gondii in mothers and risk of psychosis among adult offspring“. *Microbes Infect.* **11**: 1011–1018.

Xiong, Y., Chen, X., Chen, Z. et al. (2010). „RNA sequencing shows no dosage compensation of the active X-chromosome“. *Nat. Genet.* **42**: 1043–1049.

Yolken, R.H. and Torrey, E.F. (1995). „Viruses, schizophrenia, and bipolar disorder“. *Clin. Microbiol. Rev.* **8**: 131–145.

Yolken, R.H. and Torrey, E.F. (2008). „Are some cases of psychosis caused by microbial agents? A review of the evidence“. *Mol. Psychiatry* **13**: 470–479.

Young, S., Pfaff, D., Lewandowski, K.E. et al. (2013). „Anxiety disorder comorbidity in bipolar disorder, schizophrenia and schizoaffective disorder“. *Psychopathology* **46**: 176–185.

Zaitlen, N. and Kraft, P. (2012). „Heritability in the genome-wide association era“. *Hum. Genet.* **131**: 1655–1664.

Zheng, G. (2008). „Can the allelic test be retired from analysis of case-control association studies?“ *Ann. Hum. Genet.* **72**: 848–851.

Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J.L. (2003). „Choice of scores in trend tests for case-control studies of candidate-gene associations“. *Biom. J.* **45**: 335–348.

Zheng, G., Joo, J. and Yang, Y. (2009). „Pearson's test, trend test, and MAX are all trend tests with different types of scores“. *Ann. Hum. Genet.* **73**: 133–140.

Zill, P., Buttner, A., Eisenmenger, W. et al. (2004). „Single nucleotide polymorphism and haplotype analysis of a novel tryptophan hydroxylase isoform (TPH2) gene in suicide victims“. *Biol. Psychiatry* **56**: 581–586.

Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S. (2012). „The mystery of missing heritability: Genetic interactions create phantom heritability“. *Proc. Natl. Acad. Sci. U S A.* **109**: 1193–1198.