From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Molecular Epidemiology of Complex Heritable Disease: Applications in Genomics and Metabolomics

Robert Karlsson

Stockholm 2013

Front cover: the four panels illustrate the themes of the four papers in this thesis, while acknowledging the importance of computer-based methods in genetic research. From left to right, top to bottom, the panels show 1) deletion of genetic sequence, and characters in states of elevated and depressed mood, 2) the amino acid sequence surrounding the *HOXB13* G84E mutation (marked in black), associated with prostate cancer risk, 3) a case-parent triad and its place in the puzzle of testicular germ cell tumor genetics, and 4) identified and unidentified molecules in the human serum metabolome being investigated for their potential role in prostate cancer etiology.

Published by Karolinska institutet. Printed by Larserics Digital Print AB.

ISBN 978-91-7549-013-7

Abstract

Modern high-throughput molecular technologies (collectively referred to as “omic” platforms) are generating unprecedented amounts of data on human variation. The four papers in this thesis each investigate and characterize associations between common, complex, heritable disease, and genetic or metabolomic markers from omic platforms.

In paper I, we searched bipolar affective disorder (BPAD) pedigrees for genomic copy-number variants (CNVs, segmental deletions or duplications) segregating with disease. In one pedigree, a deletion in the gene *MAGI1* was observed in six out of six affected members. Upon further inspection, another pedigree was found with two out of three affected members carrying a duplication in the same gene. A pooled association analysis was subsequently carried out using in-house and public data sets on CNVs in control subjects and cases of BPAD, schizophrenia (SZ), or schizoaffective disorder (SA). *MAGI1* CNVs greater than 100 kb were found to be rare, nonsignificantly more common in BPAD cases than in controls, and significantly more common in the pooled case sample of BPAD, SZ, and SA than in controls.

In paper II, we studied a rare single nucleotide polymorphism (SNP) in the gene *HOXB13*, which had been recently reported to be strongly associated with prostate cancer (PC) risk. We genotyped and analyzed the variant G84E (rs138213197) in the two large Swedish PC case-control samples CAPS and Stockholm-1 (in total 4,903 cases and 4,589 controls). G84E was less rare in the Swedish samples than in the United States population previously studied, with a carrier rate over 1% in Swedish population controls. The variant was associated with a more than threefold increased relative risk of PC in both Swedish samples. G84E carriers’ absolute lifetime risk to age 80 of PC was estimated to 33%. For G84E carriers in the uppermost quartile of a genetic risk score based on common risk SNPs, the same lifetime risk was estimated to 48%.

In paper III, a replication study of previously reported genetic associations with testicular germ cell tumor (TGCT) risk was performed. SNPs in six genes (*ATF7IP*, *BAK1*, *DMRT1*, *KITLG*, *SPRY4*, and *TERT*) were genotyped and analyzed in a combined case-parent, case-control sample from Sweden and Norway. In total, 831 case-parent triads, 474 dyads, 712 singleton cases, and 3,919 control subjects were analyzed. Our results supported the previously reported association with TGCT risk for SNPs in all six genes. Tests of interaction effects revealed no allelic effect differences for the two major TGCT histological subtypes seminoma and non-seminoma. However, a variant in the gene *SPRY4* was found to differ significantly in effect depending on the sex of the parent from which it was inherited. Only maternally inherited alleles were associated with TGCT risk.

In paper IV, a large range of small molecules in human serum, collectively called the metabolome, were studied for association with PC risk and aggressiveness. Samples from 188 controls, 188 PC patients with indolent disease, and 99 PC patients with aggressive disease were analyzed by ultra-performance liquid chromatography coupled with mass spectrometry, generating 6,138 quantitative molecular features. All features were tested for association with PC status, adjusted for patient age and sample storage time. Two features were significantly associated after correction for multiple testing, but none of them could be identified as specific molecules. Testing the PC-associated features for association with 1.4 million SNPs genome-wide produced the strongest associations in variants in annotated genes, which may aid future molecular identification efforts.

In conclusion, we have used omics platforms and modern computational tools to increase our knowledge about specific genetic risk factors and metabolomic markers for complex heritable disease. Our results may come of use in future etiological research as well as in genetic and molecular risk assessment.

List of publications

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals:

I. Robert Karlsson, Lisette Graae, Magnus Lekman, Dai Wang, Reyna Favis, Tomas Axelsson, Dagmar Galter, Andrea Carmine Belin, Silvia Paddock
MAGI1 Copy Number Variation in Bipolar Affective Disorder and Schizophrenia
Biological Psychiatry. 2012; 71(10):922–930

II. Robert Karlsson, Markus Aly, Mark Clements, Lilly Zheng, Jan Adolfsson, Jianfeng Xu, Henrik Grönberg, Fredrik Wiklund
A Population-based Assessment of Germline _HOXB13_ G84E Mutation and Prostate Cancer Risk
European Urology 2012; in press http://dx.doi.org/10.1016/j.eururo.2012.07.027

III. Robert Karlsson, Kristine E. Andreassen, Wenche Kristiansen, Elin L. Aschim, Roy M. Bremnes, Olav Dahl, Sophie D. Fosså, Olbjørn Klepp, Carl W. Langberg, Arne Solberg, Steinar Tretli, Patrik K.E. Magnusson, Hans-Olov Adami, Trine B. Haugen, Tom Grotmol, Fredrik Wiklund
Investigation of six testicular germ cell tumor susceptibility genes reveals a parent-of-origin effect in _SPRY4_
Manuscript

IV. Robert Karlsson, Mun-Gwan Hong, Jessica Prenni, Corey Broeckling, Henrik Grönberg, Jonathan A. Prince, Fredrik Wiklund
Untargeted serum metabolomic profiling of prostate cancer
Manuscript

Contents

1 Introduction **1**

1.1 The central dogma of molecular biology . . . 1

1.2 It's complicated . . . 1

1.3 Classes of genetic variation . . . 2

1.4 Recombination . . . 4

1.5 Mapping disease genes in families – genetic linkage analysis . . . 4

1.6 Humanity as one big family – genetic association analysis . . . 5

1.7 Maps of the human genome – physical and genetic distance . . . 5

1.8 Heritability and heritable traits . . . 7

1.9 Omics and computing . . . 8

2 The diseases **9**

2.1 Bipolar affective disorder and schizophrenia . . . 9

2.2 Prostate cancer . . . 12

2.3 Testicular germ cell tumor . . . 14

3 Aims **17**

4 Materials **18**

4.1 The NIMH Genetics Initiative . . . 18

4.2 Case-control CNV data from publications and collaborators . . . 18

4.3 CAPS . . . 19

4.4 Stockholm-1 . . . 19

4.5 GENETEC . . . 20

4.6 TwinGene . . . 21

5 Methods **22**

5.1 In vitro . . . 22

5.1.1 Genome-wide SNP genotyping . . . 22

5.1.2 Candidate SNP genotyping . . . 22

5.1.3 Ultra-performance liquid chromatography–mass spectrometry . . . 23

5.2 In silico . . . 23

5.2.1 Linkage analysis . . . 23

5.2.2 Copy number detection from SNP chip data . . . 23

5.2.3 GWAS quality control and association testing . . . 24

5.2.4 SNP imputation . . . 25

5.2.5 Lifetime risk, adjusted for competing risks 25
5.2.6 Genetic risk scores 26
5.2.7 Combined family-based and case-control association analysis 27
5.2.8 Peak detection for UPLC-MS chromatograms 27

6 Results and Discussion 28
6.1 Rare *MAGI1* mutations increase risk for BPAD and SZ 28
6.2 The *HOXB13* G84E mutation significantly increases PC risk 30
6.3 Six genes are associated with TGCT risk, one modified by parent-of-origin . 32
6.4 Untargeted UPLC-MS metabolomic analysis of serum could not reliably separate PC patients from controls 33

7 Conclusions and future perspective 35
7.1 Implications and future research based on the papers 35
7.2 Future perspective on molecular/genetic epidemiology research 37

8 Acknowledgements 39

References 41

List of abbreviations

1kG	the 1000 genomes project
AMD	age-related macular degeneration
ATF7IP	activating transcription factor 7 interacting protein
BAF	B allele frequency
BAK1	BCL-2 antagonist/killer 1
BPAD	bipolar affective disorder
BPH	benign prostatic hyperplasia
CAPS	Cancer of the Prostate in Sweden
CI	confidence interval
cM	centiMorgan
CNV	copy-number variant
DGV	Database of Genomic Variants
DIGS	Diagnostic Interview for Genetic Studies
DMRT1	doublesex and mab-3 related transcription factor 1
DNA	deoxyribonucleic acid
ERSPC	the European Randomized Study of Screening for Prostate Cancer
FDR	false discovery rate
GWAS	genome-wide association study
kb	kilo base pairs
KITLG	KIT ligand
LD	linkage disequilibrium
LRR	log R ratio
M/Z	mass/charge ratio
NIMH	National Institute of Mental Health
OR	odds ratio
PC	prostate cancer
PLCO	the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial
PSA	prostate specific antigen
RNA	ribonucleic acid
RT	retention time
SA	schizoaffective disorder
SNP	single nucleotide polymorphism
SPRY4	sprouty homolog 4
SZ	schizophrenia
TERT	telomerase reverse transcriptase
TGCT	testicular germ cell tumor
UPLC-MS	ultra-performance liquid chromatography – mass spectrometry
WTCCC	the Wellcome Trust Case-Control Consortium

1 Introduction

This thesis is composed of four studies of genetic variation associated with human disease (papers I–III), or with metabolomic traits that are in turn associated with disease (paper IV). The first section briefly introduces key concepts of molecular biology, genetics and epidemiology upon which these studies rest, followed by a description of the heritable human diseases studied herein.

1.1 The central dogma of molecular biology

Modern molecular biology began with the discovery of the molecular structure of deoxyribonucleic acid (DNA) in 1953 [1], suggesting a mechanism for its accurate self-replication, and that the sequence of bases constituting this long, fibrous molecule may be the code that holds genetic information [2]. The hypotheses proved to hold, and since then, the molecule has been studied in relation to countless traits and diseases.

The central dogma of molecular biology (as most often described) states that information mainly flows in one direction, from DNA, through ribonucleic acid (RNA), ending up in a protein (Figure 1). This is a straightforward path from information (DNA) to function (proteins), comparable to that of a simple computer program, where information in human-readable source code format is translated by a compiler to machine code, which instructs the processor to act on data.

1.2 It's complicated

Unfortunately (to simple-minded life scientists), life is not as simple as pictured in the "central dogma", and this has been known for as long as the dogma has existed. In fact, the dogma

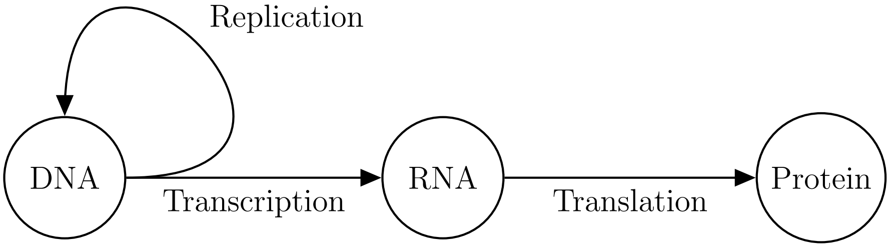

Figure 1 – The central dogma of molecular biology, as often described.

was originally stated in a negative fashion, and held that information flow from protein to DNA, from protein to RNA, and from protein to protein is *un*likely to occur (Figure 2, [3]).

While the simplified principle of “DNA to RNA to protein” is true in general, it covers far from every process in the cells of even prokaryotes (bacteria), and even less of what happens in eukaryotes (plants and animals). For example, DNA is not only a linear information carrier, but has complex structure, and is modified by epigenetic events such as methylation [4].

RNA, besides being an agent for information transfer from nuclear DNA to protein (messenger RNA, mRNA) also takes on tertiary structure to catalyze processes in the cell (ribosomal RNA, rRNA, form most of the ribosomes, which translate mRNA to protein), while small interfering RNA (siRNA) acts on complementary DNA sequences to silence specific genes [5]. Finally, long intergenic non-coding RNA (lincRNA) has been shown to take part in gene regulation and modification of the histone structure of DNA [6], and specific lincRNAs have been linked to cancer, both as tumor suppressors and as oncogenes [7].

If we not only consider information transfer, but any molecular interaction in the cell, such as enzymes acting on small and large molecules in the metabolome, proteins acting as structural support for DNA packing, et cetera, the picture becomes even more complicated, and starts approaching real life.

Returning to the computer program analogy then, real life is more like a large body of massively parallel, constantly interacting, self-modifying spaghetti code than the straightforward process suggested by the simplicity of the central dogma in its simplified form.

Nevertheless, even a vastly simplified model can be useful in making inferences and increasing understanding of the great puzzle of life piece by piece.

1.3 Classes of genetic variation

Early genetic studies used directly observable traits as indicators of which alleles had been inherited. Gregor Mendel studied, among other traits, the shape (wrinkled or smooth) and color of different varieties of peas, and especially the outcome when crossing two varieties. He used the terms *dominant*, for the traits that would always carry over to the hybrid offspring even from a single parent variety, and *recessive*, for the traits that needed to be present (possibly latent) in both parent varieties to show in the offspring [8].

In humans, early studies focused on linkage of traits to markers such as blood groups and sex, which could be easily observed and followed Mendelian inheritance rules (results included the sex-linked characters of hemophilia and red-green color blindness) [9].

After the role of DNA as the inheritance molecule and its structure was discovered, and further technical and biochemical developments, new classes of variation that could be used as genetic markers started to appear. Restriction fragment length polymorphisms (RFLPs),

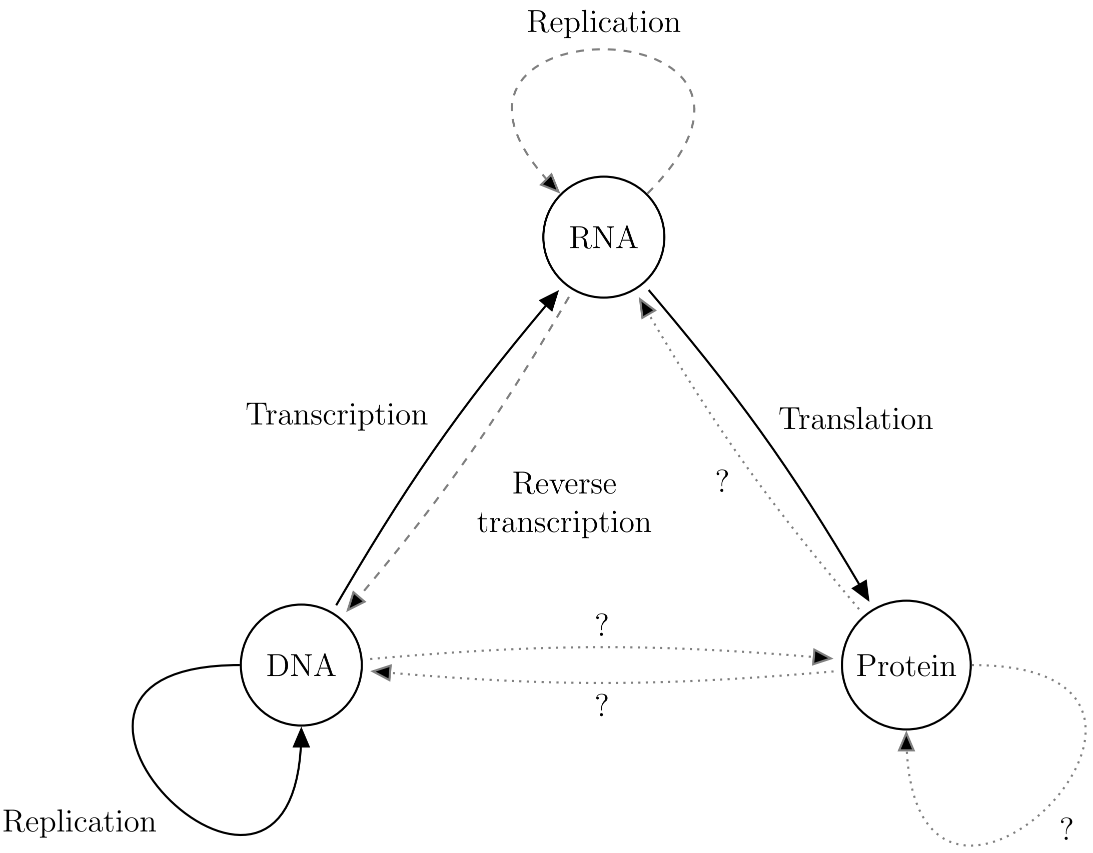

Figure 2 – The central dogma of molecular biology, as originally stated (adapted from [3]), and closer to the reality of the cell. The solid arrows indicate the general transfers, present in almost every living cell. The dashed arrows represent special transfers, which are less common and may require special conditions. Reverse transcription (RNA to DNA) is performed by reverse transcriptases in retroviruses, and in the extension of telomeres, the repetitive end regions of chromosomes. RNA replication (RNA to RNA) is performed by some RNA viruses, using special RNA replicase enzymes. Finally, the dotted arrows represent unknown transfers, which are unlikely to occur. In addition to the transfers from protein, I have included direct DNA to protein translation in this category because it has only been detected in special artificial settings [3].

variable number tandem repeats (VNTRs) and other similar polymorphisms were discovered and added to the map of the human genome, and pairwise linkage analysis of markers was used to organize them in "linkage groups", corresponding to the physical chromosomes [10]. As more and more actual genomic sequence became available, the single nucleotide polymorphism (SNP) class of mutations was found to be widespread in the genome. Although each marker provides little information, since they only have two alleles, compared to RFLPs and VNTRs which can have many, the relative abundance of SNP markers was shown to be able to compensate for the lower information content per marker. By using about three times as many markers, a biallelic SNP map of 750–1,000 markers was shown to be equivalent to the microsatellite maps of ∼300–400 markers used at the time for genome screens for linkage [11]. Today more than 50 million SNP variants are known and organized in the database dbSNP [12], where each SNP has an identifier on the form "rs", followed by a string of digits.

Even more recently, the discovery of abundant polymorphic copy-number variants (CNVs) in the human genome added another class to the list of human genetic variation [13]. CNVs are segments of the genome that have been deleted or duplicated, for example by mis-alignment of homologous chromosomes in meiosis.

1.4 Recombination

Every time a germ cell (sperm or egg) is created by cell division (meiosis), each pair of paternal and maternal homologous chromosomes recombine (cross over) at least once. This means that the germ cell chromosomes will consist of DNA sequence segments from both the paternal and the maternal homologs in the parent cell. This mechanism is the basis for both linkage and association analyses [10].

1.5 Mapping disease genes in families – genetic linkage analysis

Genetic linkage analysis is a study design which uses family data to map traits to genetic markers. The underlying hypothesis is that traits may be linked to genetic markers, meaning the trait is influenced by genetic variation a short (genetic) distance from the marker.

If a trait and a marker are linked, they tend to co-segregate in pedigrees. The amount of co-segregation under varying degrees of linkage can be modeled, and the model that fits observations the best can be estimated by maximum likelihood methods. The hypothesis of linkage versus no linkage can then be tested between a trait and genetic markers genome-wide to find the most likely linked marker (if any) [10].

1.6 Humanity as one big family – genetic association analysis

Large human populations share common ancestors if the pedigree is followed far enough back in time. Due to population history and the process of recombination, genetic markers that are physically close on a chromosome tend to be inherited together over long stretches of time. This is one of the mechanisms leading to linkage disequilibrium (LD), which is defined as a non-zero correlation between the observed alleles of two markers in a population. If a dense genomic marker map is available, and a disease-causing mutation exists somewhere in the genome, then there is a high likelihood that some of the markers will be in LD with the disease mutation, and thus correlated with disease. This is the basic concept of genetic association analysis, a study design which has dominated the field of human genetics for the past few years.

An influential paper from 1996 by Risch and Merikangas showed that for common, complex traits, where individual allelic effect sizes were expected to be modest, association studies have much better power to detect these effects than linkage analyses of comparable size. They noted that although a dense genome-wide map of SNP markers for such analysis was not currently available, this was merely a technical challenge, which would be solved over time [14]. They also suggested a genome-wide significance threshold of 5×10^{-8} to control the type I error (false positive) rate for the 1,000,000 independent tests they expected to be required for a comprehensive genome-wide association scan. This threshold is still often applied in GWAS studies of today.

An early success story of the GWAS era was a study of age-related macular degeneration (AMD). In what today would be considered a very modestly sized sample of 96 cases and 50 controls, genotyping and analysis of ~100,000 SNPs uncovered an intronic variant strongly associated with disease, and in LD with a coding, possibly causal SNP [15]. The first large (one could say massive) GWAS of common complex disease, using high density SNP chips, was published by the Wellcome Trust Case-Control Consortium (WTCCC) in 2007. In the WTCCC study, about 2,000 cases each for seven common diseases were genotyped and compared to 3,000 shared controls. 24 new risk loci were reported, but with much more modest effect sizes than those seen for AMD [16]. Today, five years later, thousands of robustly trait-associated loci have been reported from the combined effort of GWAS studies [17].

1.7 Maps of the human genome – physical and genetic distance

A number of different coordinate systems are in use for describing locations of genes and other features of the human genome. The chromosome number is common to all these systems, and is simply based on sorting the 22 pairs of autosomal chromosomes by size (see figure 3),

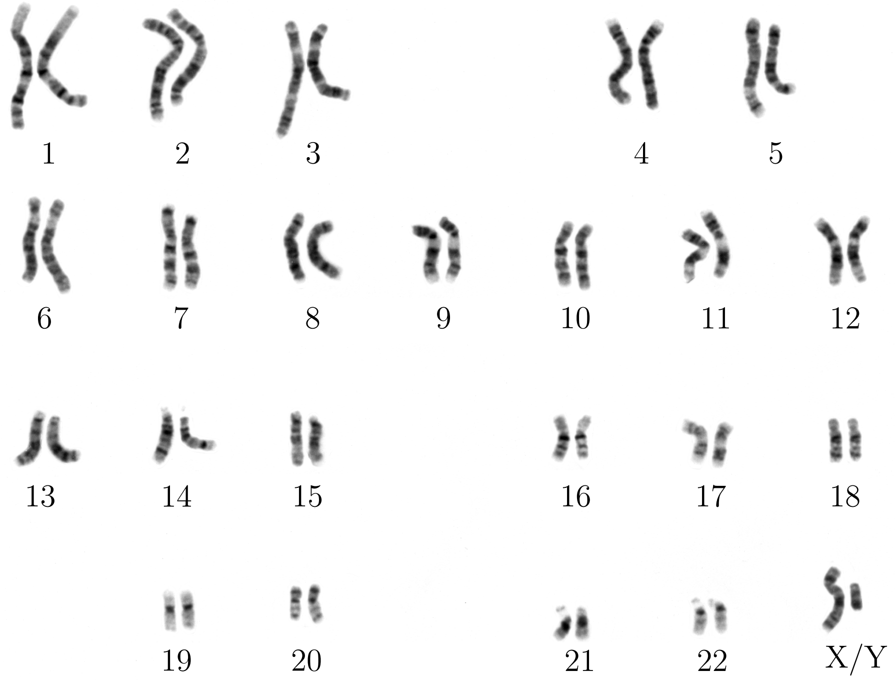

Figure 3 – Human male karyotype. Image courtesy of the National Human Genome Research Institute [18].

with chromosome 1 being the largest, and a special case for the sex chromosomes X and Y, sometimes referred to as chromosomes 23 and 24.

The most coarse of the physical coordinate systems defines genomic landmarks on the form {chr}{p/q}{band}, e.g. 8q24. The first part of the landmark identifier is the chromosome number. The next part of the identifier refers to the centromere, which is a specific region on each chromosome where it attaches to its sister chromosome during cell division. The centromere is not, despite its name, centered, but divides chromosomes into their short (petit) p arm, and the longer q arm. Thus p or q places a landmark on one of two chromosome arms. Finally, the number after the p or q identifies regions by chromosomal bands, which can be seen when stained chromosomes are examined by microscope (some bands visible in figure 3). The bands are numbered starting at the centromere and moving out along each arm. Thus, 8q24 indicates a region by the 24th band on the long arm of chromosome 8 [10].

Genetic distance, in contrast to physical distance on the DNA molecule, stems from the expected rate of recombination between two markers on the chromosome. Genetic distance is measured in centiMorgans (cM), and is defined such that the expected number of crossovers (where the germ cell's sequence switches from maternal to paternal or vice versa) in one generation between two markers that are 1 cM apart is 0.01. Genetic distance is not uniform across the physical chromosome, and furthermore differs between sexes, with the female map being longer due to a higher number of average crossovers in the production of egg cells. The length of the genetic map has been estimated to 27 Morgans in males and 39 Morgans in females [10].

Since the (almost) complete human reference sequence of approximately 3 billion base pairs became publicly available in 2001 [19, 20], a very precise physical coordinate system for genomic positions has come into common use. The system simply identifies a position by its chromosome and base pair number in the reference sequence, starting from a specified end of the chromosome. The format in text is usually “chromosome:base pair”, e.g chr8:117,700,001. Since the reference sequence is periodically updated with corrections and previously not assessed “holes in the assembly”, a version number of the reference sequence (such as hg18, hg19) must also be stated for the coordinates to be unambiguous.

1.8 Heritability and heritable traits

Heritability is a measure of the proportion of variance in phenotype (disease or trait) in a population that can be explained by genetic factors. It can be estimated by studying the phenotypic similarity of individuals of known (genotypic) relatedness, for example twin and parent–offspring pairs.

The *narrow sense* heritability is defined as the proportion of phenotypic variance in a population that can be attributed to *additive* genetic factors, with the underlying assumption of contributions coming from a large number of independent loci of small, linear, and additive effect. For qualitative traits (e.g. binary disease status), the population phenotypic variance refers to the variance of an underlying continuous liability scale, where individuals with liability over a threshold value are affected, while those with lower liability are not [21].

Traits and diseases with a high narrow sense heritability have generally been assumed to be fertile grounds for disease gene hunting. However, as results from large GWAS of common complex traits started to roll in, the variance explained by the reported associated markers was consistently lower than initially expected [22]. This phenomenon has sparked a lively debate on where the missing heritability is to be found.

A number of explanations for the discrepancy have been proposed. One line of arguments holds that variants explaining the missing heritability are there, just waiting to be found. Mechanisms supporting this standpoint are for example that most reported GWAS SNPs are most likely not the functional variant, but in LD with it, diminishing the apparent variance explained. In the same spirit, it has been suggested that rare variants, and other classes of variation such as CNVs, that are not well captured by the markers on common SNP chips, are responsible. Since GWAS methods assume that common disease is caused by common variants, rare variation may slip by unnoticed [23]. Another line of arguments addresses the heritability estimates themselves, suggesting that deviations from additivity of effects within and between loci may lead to overestimation of a trait's heritability, thereby creating "phantom heritability" which can never be completely explained by the actual disease markers [24].

Both these approaches to the missing heritability problem probably hold some truth, and future studies will most likely both reevaluate current heritability estimates, find rare variants of larger effect, and zoom in on variants previously only seen by means of LD. It would not be hugely surprising if the genetic architecture of complex disease turned out to be – complex. However, even under the simplifying assumptions underlying GWAS, important, replicable and biologically plausible findings have been made.

1.9 Omics and computing

Some of the very earliest electronic computers in the 1950s were used for linkage analysis [10]. Later, algorithmic developments (for example shotgun sequencing) and increase in computing power made possible the assembly of the first human reference genome [20]. Fruitful collaborations between biologists and mathematicians have been frequent through history – an early example being G. H. Hardy's well-known letter to Science on the expected genotype proportions in a randomly mating population under Mendelian inheritance rules (today recognized as Hardy-Weinberg equilibrium) [25].

2 The diseases

The studies in this thesis focus on four different diseases: bipolar affective disorder, schizophrenia, prostate cancer, and testicular germ cell tumor. Although seemingly disparate, these diseases have a number of features in common, which make them very interesting from a human geneticist's perspective. First, all four diseases have been shown to have a high heritability, which means that a sizable portion of disease risk is likely due to genetic factors. Second, all four diseases have a complex mode of inheritance, with no currently known major disease locus following Mendelian rules of inheritance. Finally, all four diseases considerably affect public health, and severely impact the lives of those affected.

A short description of the current state of epidemiological research, especially genetic epidemiology, for the diseases under study follows.

2.1 Bipolar affective disorder and schizophrenia

Bipolar affective disorder (BPAD) and schizophrenia (SZ) are psychiatric disorders with severe impact on the lives and wellbeing of patients and their families. Both diseases are most often diagnosed in early adulthood [26]. Though BPAD and SZ have long been regarded as separate diseases, epidemiological investigations of their co-occurrence in pedigrees [27], and high-throughput molecular genetic studies [28], have suggested that they share genetic risk factors. It has therefore been suggested that the dichotomy between the diseases is false, and that they should instead be considered as part of a wider spectrum of mood, psychosis, and autism spectrum disorders [29].

BPAD is characterized by at least one episode of mania, or mixed mania/depression (Bipolar type I), or hypomania (Bipolar type II), in combination with recurrent episodes of major depression [26]. SZ on the other hand is characterized by persistent or recurring delusions, hallucinations, disorganized speech, and/or catatonic behavior and negative symptoms such as affective flattening, alogia (poverty of speech), and avolition (an inability to participate in goal-directed activities) [26]. For a formal SZ diagnosis, the symptoms need to last for at least six months, and no mood disorder symptoms should be present. If mood symptoms (manic, depressed or mixed episodes) do occur, the patient will instead be diagnosed with schizoaffective disorder (SA).

Incidence and mortality

Both BPAD and SZ have long been thought to have a worldwide lifetime prevalence of approximately 1% each. More recent studies have updated these estimates somewhat. Systematic reviews of current incidence and prevalence studies of SZ found a median lifetime morbid risk of 0.72%, but with much variation between studies (inter-quartile range 0.47%–1.7%) [30]. Standardized mortality rates indicated a 2.6-fold higher all-cause mortality for SZ cases compared to the general population [30].

An eleven-nation study of BPAD prevalence coordinated by the World health Organization used a common diagnostic interview procedure to find cases in population-based samples from all eleven countries (Colombia, India, China, Brazil, Bulgaria, Lebanon, Mexico, Romania, Japan, New Zealand, and the United States). The overall lifetime prevalence was estimated to 0.6% for BPAD type I, and 0.4% for BPAD type II [31]. Prevalences for both subtypes varied from almost 0 in some countries up to 1% in others. These estimates are lower than those previously reported. The excess mortality for BPAD patients has been estimated in a large Swedish study to 2.5-fold in men and 2.7-fold in women (standardized all-cause mortality ratios) [32]. These figures are very similar to those for SZ, and indicate that both disorders not only lower quality of life, but can actually be lethal.

Risk factors

In the aforementioned systematic review of SZ incidence and mortality, incidence estimates differed between males and females, with a 1.4:1 rate ratio (higher in males). Furthermore, incidence was found to be higher among migrants than in native-born individuals (median rate ratio 4.6) [30]. Studies have also shown SZ incidence to be elevated in urban areas compared to rural, and in persons born in winter and spring compared to those born in the summer and fall [33].

There is little consensus on which non-genetic risk factors, if any, influence the risk of BPAD. A systematic review of around 100 previous studies found no indisputable evidence for any of a large number of investigated risk factors except for a family history of the disease, but suggested that childbirth may trigger disease onset [34]. A later study indicated that disease risk in offspring increased with increasing paternal age [35].

Genetic epidemiology

Both SZ and BPAD are often cited as highly heritable traits. A study of BPAD and SZ in the Swedish population found an offspring relative risk of 9.9 and a sibling relative risk of 9.0 for SZ, and an offspring relative risk of 6.4 and a sibling relative risk of 7.9 for BPAD. The heritability was estimated to 64% in SZ and 59% in BPAD. Interestingly, significant cross-disease risk increases were also observed, such that relatives to BPAD patients had an increased risk of SZ, and vice versa [27].

The current state of psychiatric genetics for a number of diseases, including BPAD and SZ was recently reviewed by Sullivan and colleagues [36]. Briefly summarized, the many linkage and candidate gene studies that have been performed for BPAD and SZ have rendered results that, though initially exciting, later turned out to be difficult to replicate in independent materials. This could be due to, for example, heterogeneity in the causal mutation between pedigrees and populations, lack of statistical power in the replication studies, or false positive findings in the initial studies.

A well-known early linkage study of psychiatric traits is that of a large Scottish pedigree with high prevalence of psychiatric disorders, mostly SZ, but with some pedigree members affected by BPAD or recurrent unipolar depression. A gene-disrupting translocation from chromosome 1q42.1 to 11q14.3 was found to segregate with SZ [37], and the disrupted transcripts were dubbed *DISC1* and *DISC2* for "Disrupted in Schizophrenia". Replication of this finding using association analysis of common variants in the region and a large case-control sample has been attempted, but no significant association could be seen [38]. *DISC1* variation is still being studied for association with SZ and other traits, and the validity of the initial finding has been questioned, indicating the difficulties in reaching consensus on genes' involvement in complex disease etiology [39].

Some regions (22q11, 15q13) have been shown to have recurrent CNV events (deletions or duplications) in SZ cases, and the overall genomic burden of CNV events has furthermore been shown to be higher in SZ cases [40]. A higher CNV burden in cases has also been reported in BPAD, but the finding may be limited to cases with early onset disease [36].

Results from one of the largest studies to date of SZ genetics was published by the Schizophrenia Psychiatric GWAS Consortium in 2011 [41]. SNP data from several case-control study samples were pooled into a "mega-analysis" sample of 9,394 cases and 12,462 controls, and the strongest associations from this stage were analyzed in an even larger replication sample of 8,442 cases and 21,397 controls. This resulted in seven replicating risk loci, of which five were novel. The two previously reported risk loci were located in the major histocompatibility complex region (6p21.3–22.1), and in the gene *TCF4* on 18q21.2, while the new loci mapped to the regions 1p21.3 (near sequence for the transcribed microRNA gene *MIR137*), 10q24.32 (several genes in the region), 8q21.3 (nearest gene *MMP16*), 8p23.2 (in the gene *CSMD1*), and finally 2q32.3 (nearest transcript is the non-coding RNA transcript *PCGEM1*).

A similar study of BPAD was also performed recently, coordinated by the same consortium as the SZ mega-analysis. A discovery sample of 7,481 cases and 9,250 controls was analyzed, and 34 SNPs were selected for replication in 4,496 independent cases and 42,422 controls. The study replicated the association in *CACNA1C,* first reported in a meta-analysis of the STEP-BD and the Wellcome Trust Case-Control Consortium datasets [42], and further added the *ODZ4* locus to the list of established BPAD risk genes [43].

Both the studies by the Psychiatric GWAS consortium referenced above then went on to report the results of an analysis pooling the BPAD and SZ case groups, and comparing them to a common control group. BPAD risk associations in or near the genes *CACNA1C*, *ANK3*, and the *ITIH3-ITIH4* region were strengthened when adding the SZ cases to the analysis, indicating that these could be shared susceptibility genes between the two diseases [41, 43].

2.2 Prostate cancer

Incidence and mortality

Prostate cancer (PC) is the second most common cancer diagnosis in men worldwide, and the most common in economically developed countries (using the International Agency for Research on Cancer's coarse definition from the GLOBOCAN 2008 report, "more developed regions" include all of Europe, Northern America, Australia, New Zealand, and Japan, while the "less developed regions" include all of Africa, Asia excluding Japan, Latin America and the Caribbean, Melanesia, Micronesia, and Polynesia) [44].

There is much variation in the PC incidence rates between countries, with an approximately 25-fold higher incidence in the highest-ranked region Australia/New Zealand than in the lowest-ranked region South-Central Asia [44]. These differences are most likely in part due to different underlying national rates of tumorigenesis, but also due to differences in rates of discovery in clinics, and a big part of that difference most likely due to differences in use of Prostate Specific Antigen (PSA) testing and biopsy procedures. Mortality varies less between the different regions, but an order of magnitude still separates the regions with the highest mortality rates (Caribbean) from the lowest (Eastern Asia) [44].

The situation in Sweden resembles that of many European and North American countries with an incidence that has been increasing rapidly starting in the 1990s, following the introduction of the PSA test. Meanwhile, PC mortality has been more or less constant over the past half century (Figure 4, data from NORDCAN [45]).

Prostate specific antigen testing

The prostate-expressed glycoprotein PSA was shown to be a sensitive serum biomarker for PC diagnosis and progression in 1987 [46]. The test lacks in specificity, since patients with benign prostate hyperplasia (BPH) also have elevated serum PSA levels. Nevertheless, PSA testing has since become immensely popular, and much of the increased PC incidence in developed countries may be attributable to increased PSA testing [47]. Despite there being no formal PSA screening program in Sweden, the current extent of PSA testing can be seen as an ad hoc and uncontrolled screening program. A recent study of the prevalence of PSA

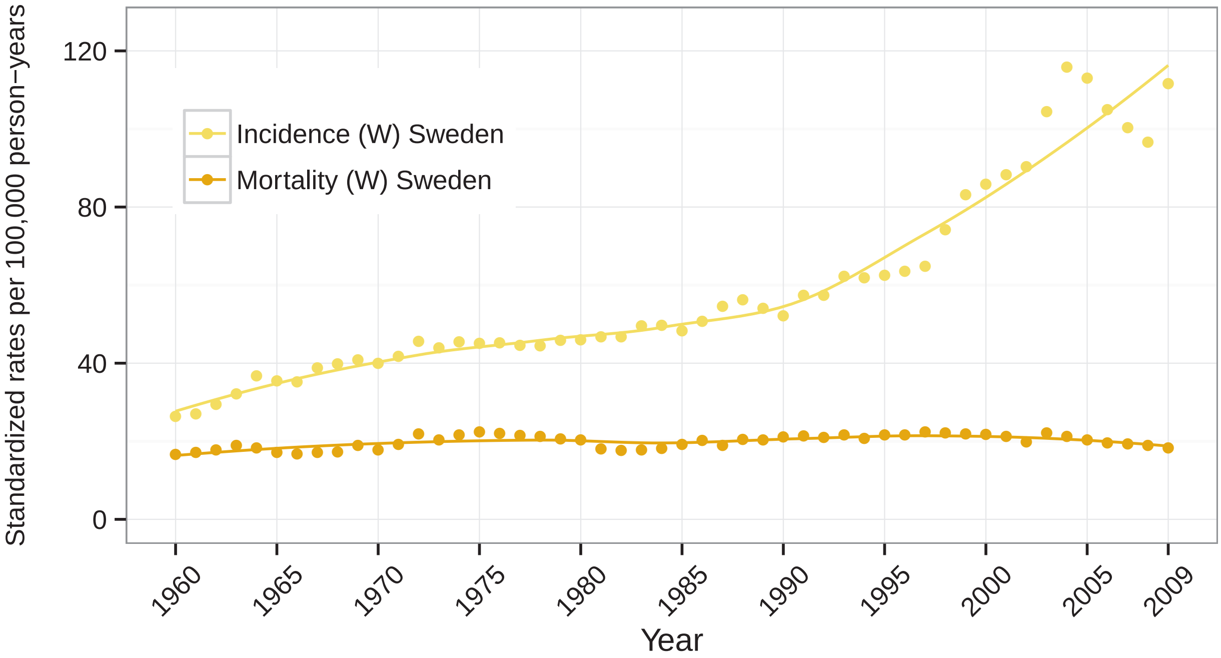

Figure 4 – Prostate cancer incidence and mortality in Sweden 1960–2009. Rates are the number of new cases/deaths per 100,000 person-years, standardized to the World Standard Population. Points represent actual reported rates, while lines are LOESS smoothed estimates. Data from NORDCAN [45].

testing in Stockholm county reported that in 2011, 46%, 68%, and 77% of men in the age groups 50–59, 60–69, and 70–79 years respectively had performed at least one PSA test in the last 9 years [48].

Two large randomized trials of the impact of organized PSA testing on PC mortality and overall mortality, the European Randomized Study of Screening for Prostate Cancer (ERSPC) [49] and the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) [50] studies, have recently published results from 11 and 13 years of follow-up respectively. Results were conflicting, with ERSPC reporting a statistically significant improvement of PC mortality in the screening group, while PLCO reported no significant improvement. The difference between the results may be due to a higher degree of opportunistic screening in the control arm of PLCO, or to a number of other differences between study designs and the underlying populations. The dispute on whether organized PSA screening is advisable is thus far from settled.

Risk factors

Besides genetic risk factors including family history, which will be discussed separately, there are a number of established and suspected risk factors for PC. The established risk factors include ethnicity (e.g. low incidence in Asian populations, high in Northern Europe, and very high in African-Americans in the USA) and age (almost no one is diagnosed with PC before age 45, but the incidence increases rapidly with age after that) [51]. Dietary risk factors for PC have been investigated in many studies, but results are not yet conclusive. Observational studies have among other things suggested increased PC risks from high dietary fat and red meat intake [52], and protective effects from fatty fish [53], phytoestrogens [54], vitamin E, selenium, tomatoes and lycopene (an antioxidant found in high levels in tomatoes), cruciferous vegetables (such as cabbage, broccoli, cauliflower, and brussels sprouts), and green tea [52]. Some of these effects have support from in vitro studies, but the evidence from large randomized clinical trials for chemoprevention has so far not been convincing [52].

Genetic epidemiology

Prostate cancer is one of the most heritable forms of cancer. Studies of twins have estimated the proportion of variance in PC risk that is due to genetic factors to 42%, and a family history of prostate cancer is one of the strongest known risk factors for the disease [55, 51].

Linkage and candidate gene studies have not resulted in many replicated findings, symptomatic of a disease with complex inheritance patterns. Some genetic variants in the loci *BRCA1* and *BRCA2* strongly associated with breast cancer risk have also shown association to prostate cancer risk [56].

Great advances in the search for PC susceptibility loci have been made with the feasibility of genome-wide association studies. GWASs and GWAS meta-analyses have uncovered several loci consistently associated with PC risk during the last few years, albeit most with small effect sizes. The current landscape of established genetic associations with prostate cancer has been recently reviewed by Goh and colleagues [57], and consists mostly of common variants of small effect, derived from GWAS analyses.

Since the review by Goh et al was published, a newly discovered relatively rare single nucleotide variant in the gene *HOXB13* has been associated with a high relative risk of prostate cancer [58]. This association was replicated in a large Swedish case-control sample in paper II of this thesis, and will be discussed further below.

2.3 Testicular germ cell tumor

Incidence and mortality

Most malignant neoplasias (approximately 95%) in the testes are germ cell tumors [59]. Although testicular germ cell tumor (TGCT) is an overall rare disease, with approximately 6.2 new cases per 100,000 person-years in Sweden (average between years 2000 to 2009, data

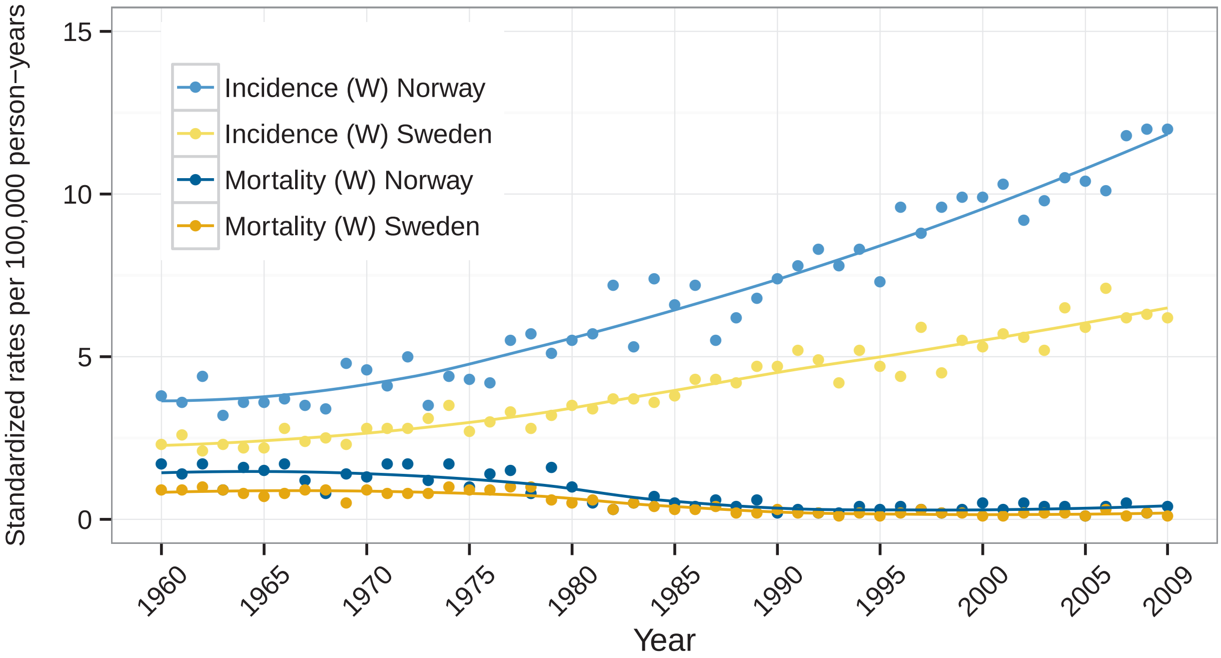

Figure 5 – Testicular cancer incidence and mortality in Sweden and Norway 1960–2009. Rates are the number of new cases/deaths per 100,000 person-years, standardized to the World Standard Population. Points represent actual reported rates, while lines are LOESS smoothed estimates. Data from NORDCAN [45].

from NORDCAN [45]), its incidence has been increasing over time worldwide, nearly doubling since the 1960s [59]. Furthermore, it is the most commonly occurring solid tumor in young men (between approximately 15–35 years of age) in the Nordic countries [45], with similar patterns seen for North America, Western Europe, and Australia [60]. Figure 5 shows the incidence and mortality of TGCT since 1960 in Sweden and Norway, age-standardized to the World Standard Population.

Due to an uncommonly high sensitivity of TGCT cells to combination chemotherapy including cisplatin, most patients since the mid-1980s (when the treatment regimen was introduced) survive their disease. This holds true even for late stage and metastatic disease [59]. The 5-year relative survival (with 95% confidence intervals) was 94% (89–98) in Sweden and 90% (86–96) in Norway for testicular cancers diagnosed between 1999 and 2008 [45].

Risk factors

Even though most TGCTs are thus curable, it is still of interest to understand the underlying etiology, and if and how they could have been prevented. Besides the immediate use of newly discovered risk factors for risk assessment, some insights in the disease mechanisms could translate to other, more lethal cancers, and increase our understanding of their etiology, and in the very long run provide clues on how to cure them. To this end, several studies of the epidemiology of TGCT have been performed.

Among the established or suggested risk factors for TGCT we find undescended testes (cryptorchidism), a previous (contralateral) TGCT, hormone exposures *in utero*, perinatal factors, and a family history of the disease [61]. Male subfertility has also been associated with increased TGCT risk, an association which remained when excluding subfertility due to cryptorchidism [62].

Genetic epidemiology

First-degree relatives of TGCT patients have a highly increased disease risk, with one large study of the Swedish population reporting a standardized incidence ratio of 3.8 (95% CI 2.2–6.2) for sons of cases and 8.6 (95% CI 6.4–11.3) for brothers of cases [61]. The heritability of TGCT has been estimated in family studies to 25% (95% CI 15–37), which is one of the highest heritabilities among cancer diagnoses [63]. This indicates that genetic factors play an important role in the disease etiology.

Several genetic associations have been reported for TGCT risk. The 1.6 mega base microdeletion gr/gr on the Y chromosome was found in year 2005 to be associated with a two- to threefold increase in TGCT risk [64]. More recent findings from GWAS efforts include associations within the genes activating transcription factor 7 interacting protein (*ATF7IP*), BCL-2 antagonist/killer 1 (*BAK1*), doublesex and mab-3 related transcription factor 1 (*DMRT1*), KIT ligand (*KITLG*), sprouty homolog 4 (*SPRY4*), and telomerase reverse transcriptase (*TERT*) [65, 66, 67, 68]. A meta-analyses of current GWAS efforts, results of which were presented at the 2012 annual meeting of the American Society for Human Genetics, implicated four additional loci associated with TGCT risk [69]. These were located in chromosomal regions 4q22.2 (in the gene *HPGDS*), 7p22.3 (in the gene *MAD1L1*), 16q22.3 (in the gene *RFWD3*), and 17q22 (in or near genes *TEX14*, *PPM1E*, and *RAD51C*). These results are however to be considered as preliminary, since they have not yet been published in a peer-reviewed journal.

3 Aims

The general aim of this thesis was to increase knowledge of disease-related genetic and metabolomic variation in heritable complex diseases and disorders, by applying appropriate computational and statistical methods to large datasets in a molecular epidemiological framework.

The specific aims of this thesis, each corresponding to one of its four component papers, were:

- To find rare, highly penetrant, disease-associated genetic variants in bipolar disorder, through a comprehensive assessment of genetic variation in families with high prevalence of the disease, and to follow up these findings in a larger case–control sample of bipolar disorder and schizophrenia.

- To assess the prevalence and penetrance of a newly discovered rare genetic variant, associated with a high risk for developing prostate cancer, in two large Swedish case-control samples, and to further assess its impact on lifetime risk of the disease.

- To replicate and further characterize findings from genome-wide association studies of testicular germ cell tumor, through genotyping and analyzing tagging SNPs in case-parent triads and unrelated cases and controls.

- To study whether metabolites in serum, detectable trough ultra-performance liquid chromatography–mass spectrometry (UPLC-MS), can be used as biomarkers, in order to separate prostate cancer cases from healthy controls, and indolent prostate cancer from more aggressive disease.

4 Materials

Patient materials for the papers in thesis were assembled from multiple large case-control and family-based study samples.

4.1 The NIMH Genetics Initiative

The National Institute of Mental Health (NIMH) Bipolar Disorder Genetics Initiative [70] has established a repository of data and biomaterials (DNA and lymphoblastoid cell lines) from pedigrees with at least one proband affected by BPAD, with resources available upon request for qualified researchers investigating BPAD genetics. Data collection began in 1991, and is still ongoing. From 1996 and onwards, the project has been accepting data requests.

In addition to the biological samples and cell lines, the project database includes anonymous data on family structure, age, sex, vital status, psychopathology, diagnosis, and other clinical information, as acquired through relatives, medical records, and directly through the Diagnostic Interview for Genetic Studies (DIGS) structured interview tool.

For paper I, a data request was granted for pedigrees with available genome-wide microsatellite data from previous linkage analyses. Previous reports had only published compound linkage results from entire waves of pedigrees. In order to select the most promising pedigrees for rare variant discovery, we therefore performed family-wise linkage analyses using the available markers. Based on these analyses, a request for DNA from 277 individuals in 48 pedigrees was made and granted.

Genome-wide SNP genotypes for 592,275 markers were then successfully generated for 275 of the DNA samples using the Illumina Human 610 quad chip, and analyzed further in house.

4.2 Case-control CNV data from publications and collaborators

Stage 2 of the analyses in paper I involved a large, pooled sample of CNV genotypes for cases affected by BPAD, SZ, or SA, and control samples. These genotypes were assembled partly from samples available to collaborators, partly from complete datasets published as supplementary information to original articles, and partly from direct contact with authors of studies on copy-number variation in psychiatric or other disease. A full listing of the sample sources can be found in table 1 of paper I. Since some of the included studies reused the same samples, care was taken when assembling the data to only count each individual once. In total, we assembled data from 3,683 BPAD cases, 7,242 SZ or SA cases, and 16,747 controls.

4.3 CAPS

Cancer of the Prostate in Sweden (CAPS) is a population-based case-control study of genetic and dietary PC risk factors. Patients were identified from regional cancer registries between 2001 and 2003. Control subjects were recruited concurrently using the Swedish population registry. Controls were selected, randomly from the total male population, to match the expected age (in five-year age groups) and geographic distribution of the cases.

Information and biomaterials collected for each participant included a blood sample, and a questionnaire with questions concerning family history of PC, diet, and lifestyle. For the cases, data also include date of diagnosis, diagnostic PSA measurement, and tumor characteristics. Blood samples from cases were taken on average five months after the time of diagnosis.

The sample database has been continuously updated with genotypes of new potential risk markers for PC over the years. One of the most recent updates provided genotypes for a rare coding variant (and surrounding common variants) in the gene *HOXB13*, rs138213197, which is further discussed in paper II.

Recently genome-wide SNP genotypes from the Affymetrix 500k and 5.0 chips were added to the body of data available for the sample. These genotypes have been successfully used in meta-analyses of PC risk and aggressiveness [71, 72], and were studied in paper IV of this thesis for their relation to serum metabolomic features associated with PC. Finally, as previously mentioned, a subsample consisting of controls (188), cases with indolent disease (188), and cases with aggressive disease (99), was selected for analysis of serum metabolomics using UPLC-MS, which is further described in paper IV. The numbers of cases and controls with the different types of molecular data available are presented in table 1.

Available data	Cases	Controls	Used in paper
Blood sample, questionnaire and registry data	3,161	2,149	
HOXB13 G84E genotypes	2,805	1,709	II
Genome-wide SNP genotypes	1,932	994	IV
UPLC-MS spectrograms	287	188	IV

Table 1 – The CAPS study sample of PC patients and population controls.

4.4 Stockholm-1

Stockholm-1 is a study of men in Stockholm, Sweden, who underwent a prostate biopsy examination between 2005 to 2007.

Patients were identified through patient registries, and invited to participate in the study by providing a blood sample and filling in a questionnaire regarding family history of PC. In total 5,241 men (2,135 biopsy positive cases and 3,106 biopsy negative controls) consented to participate and provide the required materials. In addition to the data provided directly by the patients, the study includes information on PSA levels as measured before the biopsy was taken (obtained by registry linkage to the PSA testing laboratories in Stockholm), biopsy results from linkage to the Pathology laboratory, and cancer status through linkage to the regional cancer and prostate cancer quality registry.

The sample was first used by Aly and colleagues to show that adding a score based on 35 PC risk SNPs to a risk prediction model based on PSA, age, and family history of PC could increase the specificity of the prediction model without reducing sensitivity [73].

For paper II in this thesis, the rare coding variant rs138213197 (G84E) in the gene *HOXB13* was genotyped with a number of surrounding common SNP markers in order to assess their impact on PC risk. For this study, patients who had displayed biopsy results positive for PC were designated as cases, while the biopsy negative men were used as controls. 2,098 cases and 2,880 controls were successfully genotyped and analyzed. Furthermore, genotypes for the established risk SNPs studied by Aly and colleagues were available and investigated for their ability to predict lifetime PC risk in combination with the moderately penetrant *HOXB13* variant.

4.5 GENETEC

The GENETEC study sample of TGCT patients and their parents was collected in Sweden and Norway between 2008 and 2010. Eligible patients were identified through national cancer and patient registries. All men with a TGCT (ICD-10, C62) diagnosis between 1995 and 2006 in Sweden and between 1990 and 2008 in Norway were invited by mail to participate in the study. Patients who chose to participate were asked for permission for us to contact their biological parents for invitation to participate in the study. DNA was collected from saliva. Both patients and their parents were sent the same DNA self-collection kit with instructions to “spit in the tube” and return the sample by prepaid mail.

The number of participating TGCT patients and parents per country are displayed in table 2. Even though Sweden has almost twice the population of Norway, the higher rate of

	Norway		Sweden		Total	
TGCT cases	974		1,188		2,162	
Full triads	483		521		1,004	
Dyads (mothers/fathers)	192	(150/42)	248	(178/70)	440	(328/112)
Singletons	299		419		718	

Table 2 – The GENETEC study sample of TGCT patients and their parents.

TGCT in Norway led to a sample that was almost balanced between the countries. Average age at diagnosis among TGCT cases was 32 years (range: 15 to 65 years).

Kristiansen and colleagues published the first results from this study sample, in a genetic association study of SNP variation in candidate genes in sex hormone pathways in relation to TGCT [74].

4.6 TwinGene

The TwinGene study sample is a population-based sample of Swedish twins (monozygous and dizygous) identified from the Swedish twin registry, and ascertained between 2004 and 2007. Study participants provided information about zygosity, lifestyle and health by questionnaire, and a blood sample for extraction of DNA and other blood components. The total sample size was 12,591 individuals, with year of birth ranging from 1911 to 1958 [75]. A majority of the sample (9,836 individuals) were genotyped genome-wide using the Illumina Human OmniExpress bead chip, providing genotype information for about 730,000 SNP markers per individual.

For paper III of this thesis, a subsample of the TwinGene cohort was selected to act as population controls in the case-control analysis of TGCT risk SNPs. We created this control group by extracting all unrelated male subjects among the samples that were successfully genotyped. Thus, if a twin pair consisted of a brother and a sister, we included the brother. If the twin pair was all male, one of the brothers was extracted (assuming both had been successfully genotyped). Through this selection procedure, 3,919 control samples were included in the analyses for paper III. As for the case-parent sample, this control group was also used in the study of genes in hormone pathways by Kristiansen et al. [74].

5 Methods

5.1 In vitro

This section describes the biochemical methods that were used to generate data from biological samples.

5.1.1 Genome-wide SNP genotyping

In the early days of SNP genotyping, the process was manual and tedious. Technical advances quickly improved the procedure, and more and more markers could be genotyped in parallel in ever shorter timespans. The SNP genotypes used for association testing in the TwinGene sample in paper III, the CAPS sample in study IV, and for assessment of CNVs in paper I, were all measured using genome-wide SNP chips, a technology which enabled the genome-wide association study (GWAS) era.

Based on the human reference sequence, and the database of known SNP locations, short primer DNA sequences are synthesized and placed on specific locations on a small chip. Modern SNP chips can measure genotypes for up to a few million markers in parallel. Fragmented sample DNA is then hybridized to the primers, and based on either allele-specific base extension or allele-specific binding, an optical signal is generated. The signals are digitized by a specialized scanner instrument, and computer algorithms use the intensity and color of each signal to determine individual genotypes. A comprehensive review of the technologies underlying current GWAS genotyping platforms can be found in [76].

5.1.2 Candidate SNP genotyping

When only a small number of specific SNPs are of interest, genotyping markers across the entire genome on a SNP chip is a waste of time and resources. Before the GWAS era, genotyping a few markers at a time was the only way to acquire genetic data. Today, similar but much optimized methods are used as a quick and non-expensive complement to genome-wide SNP chips. The main uses today of these “oligoplex” technologies are replication studies of GWAS findings, candidate gene studies, and clinical genotyping.

The Sequenom iPLEX MassArray technology was used in papers II and III of this thesis for replication of previously published genetic associations. The platform uses allele-specific base extension in the genotyping process, but instead of optical signaling, alleles are detected by mass spectrometry [76].

5.1.3 Ultra-performance liquid chromatography–mass spectrometry

In order to measure the abundances of a large slice of the spectrum of molecules present in human serum (human blood with cells and coagulant factors removed), we applied ultra-performance liquid chromatography coupled with mass spectrometry (UPLC-MS) for the analyses in paper IV.

The UPLC-MS technique is a method which spreads molecules in a sample along two dimensions. The first step (UPLC) separates a sample by a gradient of molecule size and solubility in a solvent fluid, pumped through a column. This generates a retention time (RT) axis (different molecules take varying amounts of time to move through the UPLC column due to differences in size and solubility).

At the end of the UPLC column, the sample moves to the mass spectrometry step, where an electrical voltage is applied to fling molecules across a small gap, and smash them into an ion detector. The molecules land on different locations on the detector depending on the relation of their molecular mass to their electrical charge, the mass/charge ratio (M/Z). For each molecule eluting through the column and being measured by the ion detector, an intensity which is proportional to the abundance of the molecule in the serum is recorded, along with its M/Z and RT values.

The resulting three-dimensional data matrix is further processed to transform data from the continuous M/Z and RT spectrum to a list of "peaks" corresponding to detected ions, and their intensity for each serum sample analyzed. This process is described under the following *in silico* section.

5.2 In silico

This section briefly describes the statistical, epidemiological, and computational methods which were used to manage, analyze, visualize, and interpret the data.

5.2.1 Linkage analysis

In paper I, family-wise linkage analysis using microsatellite and SNP markers (for the pedigrees genotyped on SNP chips) was performed using GENEHUNTER-PLUS [77]. Recessive, dominant, and nonparametric models were used.

5.2.2 Copy number detection from SNP chip data

For generation of genome-wide CNV data from probe intensity data from genome-wide SNP chips, we used the software package PennCNV [78]. PennCNV uses a hidden Markov model to detect genomic copy number state from the absolute intensity (log R ratio, LRR) and relative

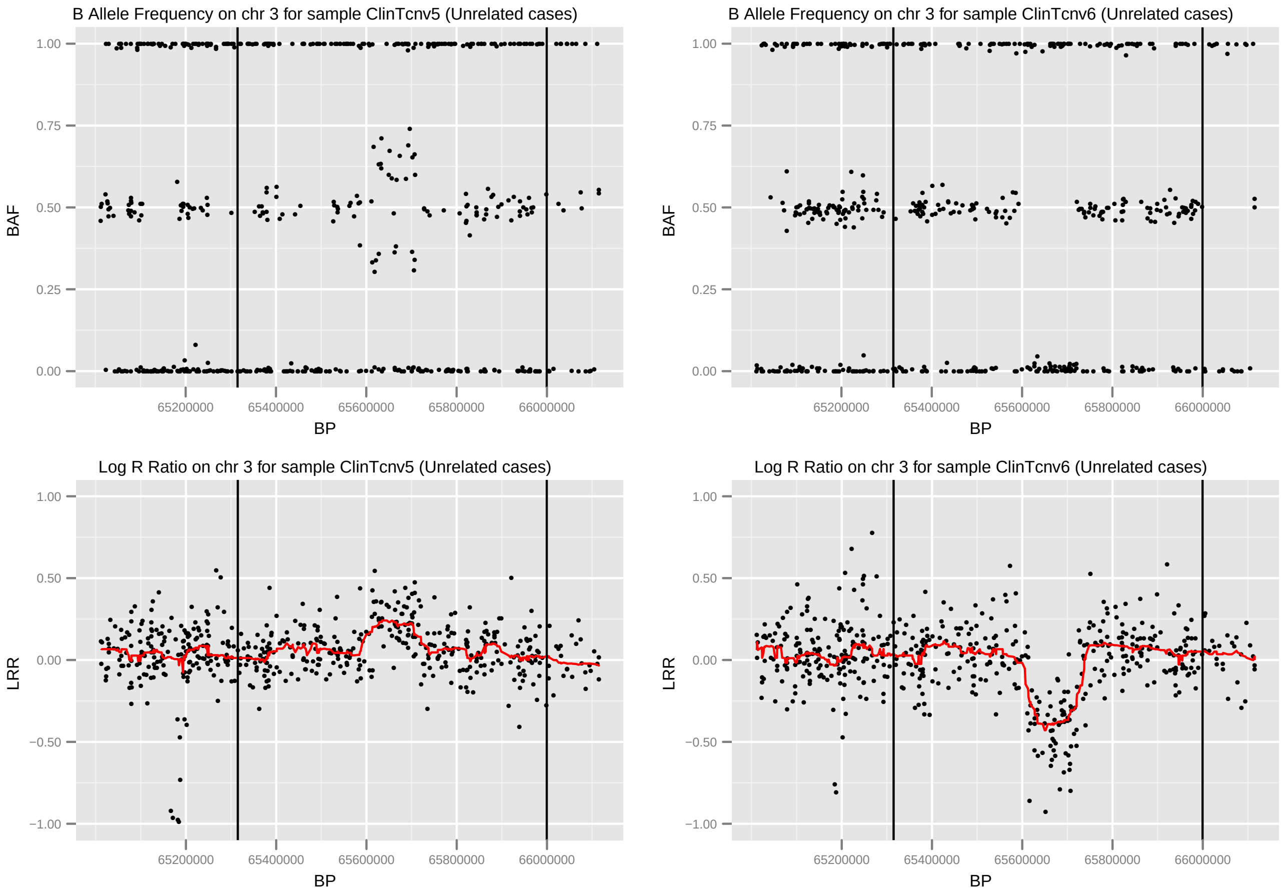

Figure 6 – SNP data for CNV detection by PennCNV. The left column shows a duplicated region (copy number is 3). The upper BAF panel shows clusters for the four genotypes AAA, AAB, ABB, and BBB for a chromosomal segment where the lower LRR (intensity) panel shows an increase in signal strength. In the right column a deletion event has occurred (copy number is 1). The upper BAF panel shows clusters for the two genotypes A and B for a chromosomal segment where the lower LRR panel shows a decrease in signal strength.

intensity of allelic probes (B allele frequency, BAF) sequentially along the chromosomes. Figure 6 displays example raw data for a duplication and a deletion as detected by PennCNV.

5.2.3 GWAS quality control and association testing

Because of the large number of tests performed in a standard GWAS, meticulous quality control of data is essential for controlling the number of false positive findings. Standard steps of quality control include the exclusion of individuals and markers with a high rate of missing genotypes, tests for deviations from Hardy–Weinberg equilibrium, which may indicate genotyping problems, and the assessment of and adjustment for population substructure in the sample, which may inflate false positive rates. Standard protocols have been developed for this quality control procedure [79], but must often be adapted to be of use to any given research group.

It has been said that standard GWAS data quality control is a sort of "post mortem" rescue attempt of a failed experimental design, and that more care should be taken to properly randomize case and control samples with regard to analysis batch parameters [80]. Doing this would make many standard data exclusion steps unnecessary, because the case-control analysis is not confounded by randomized factors. However, because GWAS data sets are often re-used for analyses not originally planned, pooled in meta- and mega-analyses, and because a large number of GWAS data sets have already been genotyped, the post-genotyping quality control process will be in use for a long time.

Standard tools for quality control and analysis of GWAS data include the PLINK [81] software toolbox, and R [82].

For the association analyses in paper II, we used logistic regression modeling in R to calculate odds ratios (OR), approximating relative risks, and 95% confidence intervals (CI). For the haplotype analyses in the same paper, we used PLINK. Data quality control in paper III used R and PLINK, which is also useful when managing and cleaning candidate gene data. Finally, for the genetic association step of study IV, PLINK was used to perform linear regression analysis of metabolomic features and SNP markers.

5.2.4 SNP imputation

In order to artificially increase the marker density of the candidate gene regions of paper III, we performed genotype imputation using the Beagle software package [83, 84]. Beagle uses information on family structure and local LD together with a densely genotyped reference panel (1000 genomes) to infer markers which were not directly genotyped with good accuracy.

5.2.5 Lifetime risk, adjusted for competing risks

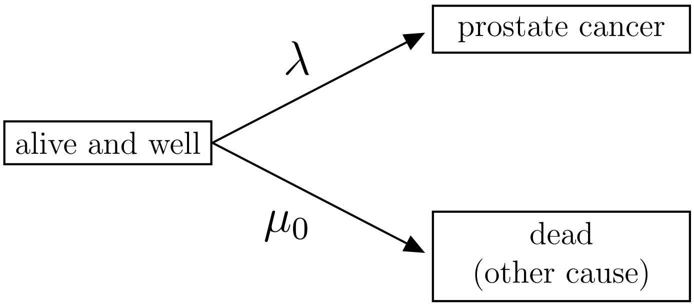

Figure 7 – State diagram for the competing risk model used in paper II. μ_0 and λ are the transition rates between the different states: μ_0 is the other cause mortality rate, and λ is the prostate cancer incidence rate.

In paper II, we estimated the effects of the *HOXB13* G84E mutation on lifetime absolute risk of prostate cancer in Swedish men. In order for the estimates to be applicable to the real world, we needed to take into account the competing risk of other cause mortality. This is especially important for a disease with late age of onset such as PC, since other cause mortality is considerable in older age groups.

We first used the Swedish total population registries of cancer incidence and causes of death (Statistics Sweden) to acquire the national rates of PC incidence and other cause mortality (the latter by subtracting the number of PC deaths from the all-cause mortality).

With population rates of PC incidence and other cause mortality available, we estimated the cumulative hazard of a PC diagnosis from birth to age t years, adjusting for competing risks (other cause mortality) by numerically solving the differential equation system

$$\begin{cases} s'(t) = -(\lambda + \mu_0)s(t) \\ c'(t) = \lambda s(t) \\ s(0) = 1,\ c(0) = 0 \end{cases}$$

for $c(t)$, where $s(t)$ is the cumulative survival from age 0 to age t, $c(t)$ is the cumulative PC incidence from age 0 to age t, λ is the group specific incidence rate for G84E carriers and non-carriers (estimated from the logistic regression odds ratios, adjusted for the sampling age strata, and the population incidence), and μ_0 is the other cause mortality. The possible states and transitions of persons in the analysis is displayed in figure 7. Proportional hazards between *HOXB13* G84E carriers and non-carriers were assumed.

5.2.6 Genetic risk scores

In paper II, we combined the moderately penetrant *HOXB13* G84E mutation in a statistical model with a genetic risk score (also called polygenic risk score) based on known risk loci of small effect. Such scores were used in an innovative manner to show that many risk loci for SZ of very small effect are, in compound as a polygenic risk score, associated with BPAD [28]. They have also been used in other efforts including PC risk prediction [73].

Some studies using the risk score approach have simply created the score as a count of the number of risk alleles carried by an individual, while others have weighted the count at each marker locus by the per-allele odds ratio, calculated from the same material or from literature surveys or other samples. The approach we used in paper II was to search the literature for risk markers, calculate their allelic odds ratios in the Stockholm-1 sample, and then generate the per-individual scores in the CAPS sample using the Stockholm-1 odds ratios as weights.

In mathematical notation, our score was defined as

$$S_j = \frac{1}{n} \sum_{i=1}^{n} \log(\mathrm{OR}_i) a_{ij},$$

where S_j is the risk score for individual j, OR_i is the per allele odds ratio for marker i (out of a total of n markers), and a_{ij} is the number of non-reference alleles (0, 1, or 2) carried by individual j at marker i. This formula assumes that the risk contributions from each SNP combine in a multiplicative manner.

5.2.7 Combined family-based and case-control association analysis

For the association testing of paper III, we applied a likelihood-based test able to combine information from case-parent units with case singletons and unrelated control subjects in a single powerful test for association. The test is implemented in the software package UNPHASED [85].

5.2.8 Peak detection for UPLC-MS chromatograms

For the conversion of raw UPLC-MS data to a list of molecular features in paper IV, we used the software package XCMS [86]. After normalization, averaging intensities over sample triplicates, and log10-transformation of the detected features, association analyses were performed by linear regression, and the ANOVA F-test. Peak detection, normalization, and analysis was performed in the R statistical programming environment [82].

6 Results and Discussion

6.1 Rare *MAGI1* mutations increase risk for BPAD and SZ

In paper I we describe a search in 48 BPAD pedigrees (277 individuals) for rare CNVs segregating with disease. After CNV detection, variants shorter than 10 kilo base pairs (kb), and common variants described in the public Database of Genomic Variants (DGV) were filtered out. After filtering, CNVs were ranked by the number of affected individuals per family in which they were detected. A CNV of seemingly high penetrance in a large pedigree (an intronic deletion of ∼200 kb) was found in the gene *MAGI1*, and this genomic region was consequently studied further in a large pooled sample of unrelated cases and controls.

Since multiple lines of evidence have indicated that BPAD, SZ and SA may share genetic risk factors, the decision was made for the case-control replication effort to study cases from not only BPAD samples, but also those affected by SZ and SA. If the CNVs we investigated were among the shared genetic risk factors, this approach would then increase the statistical power of this study. In total, the pooled dataset consisted of 3,683 BPAD cases, or 10,925 cases when pooling the samples of BPAD, SZ, and SA patients, and 16,747 control samples. The data sources are described briefly in section 4.2, and fully in Table 1 of paper I. Since many sources only reported variants of 100kb or larger, and the detected variants in our family sample were well above that threshold, the pooled analysis was limited to variants exceeding this size limit.

One additional CNV in the region, a duplication of ∼160 kb, was found in two out of three affected members of another BPAD pedigree in the family sample. We counted this as one exposed case for the case-control analysis. A further event was detected in a BPAD case from the Wellcome Trust Case-Control Consortium sample. The pedigree structures of the two BPAD families with *MAGI1* CNVs are displayed in figure 8. When expanding the case sample to include SZ and SA cases, five additional CNVs were detected in cases from

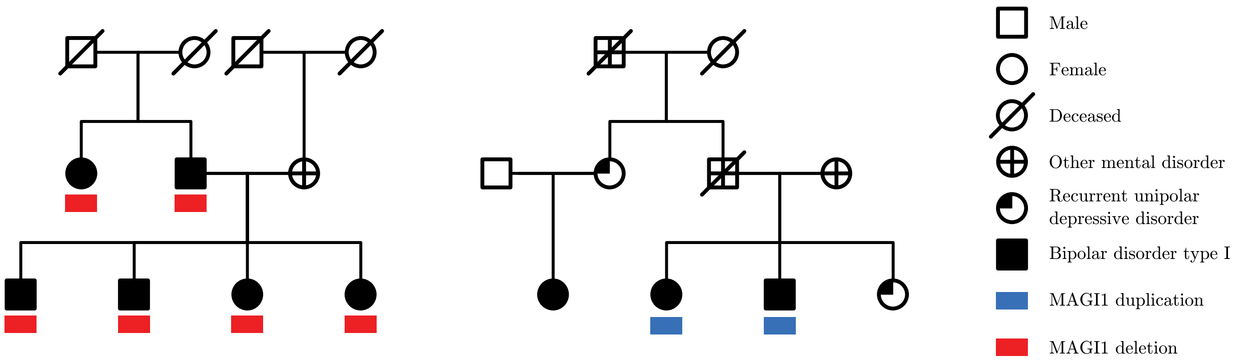

Figure 8 – BPAD pedigrees in which *MAGI1* CNVs were detected.

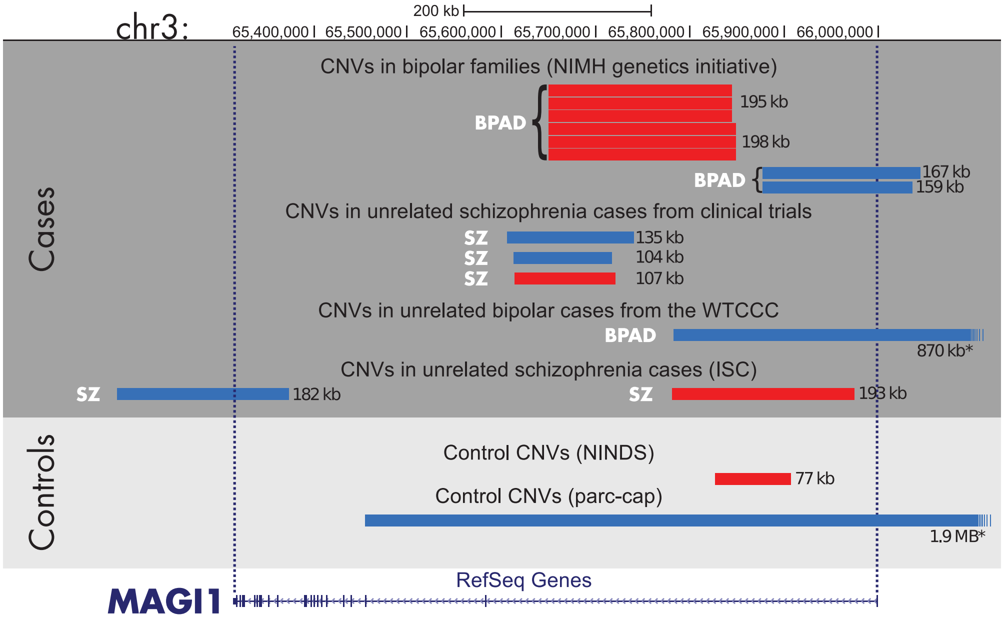

Figure 9 – *MAGI1* copy-number variation in BPAD pedigrees and BPAD, SZ, SA, and control samples of unrelated individuals. Red bars represent deletions, blue bars duplications. White text in the "Cases" section mark diagnoses in carriers. The bottom line shows the extent of the coding (wide line) and non-coding (narrow line with arrowheads) genomic sequence of the *MAGI1* gene. Genomic coordinates refer to the hg18 reference assembly of the human genome.

the International Schizophrenia Consortium and clinical trial participants from Johnson & Johnson Pharmaceutical Research and Development.

In all of the 16,747 pooled control samples, only two CNV events affecting the MAGI1 region were detected. Figure 9 displays the genomic location of all the *MAGI1* CNVs used for association testing in this study.

CNV-disease association was tested in the pooled case-control sample using Fisher's exact test, with a one sided alternative hypothesis due to the prior hypothesis of *MAGI1* CNVs being more common in cases. The pedigree in which the variant segregating with BPAD was first discovered was not included in any test for association, and the duplication appearing twice in another pedigree was only counted once. The results for the two analyses of BPAD cases versus controls, and BPAD, SZ, and SA versus controls are presented in table 3 (reproduced from paper I).

MAGI1 CNV events were significantly more common in the pooled BPAD, SZ, and SA case group than in the control population. Due to the rarity of these events, effect estimates could not be made with high precision, reflected by the very wide confidence intervals presented. Even larger study samples are needed in order to assess the true impact of these variants. Nevertheless, the present study strongly supports the importance of the *MAGI1* gene in BPAD etiology.

Hypothesis	Cases	Controls	Case CNVs	Control CNVs	P	OR (95% CI)
MAGI1 CNVs more common in cases						
BPAD alone	3,683	16,747	2	2	0.15	4.5 (0.5–∞)
BPAD, SZ, and SA	10,925	16,747	7	2	0.023	5.4 (1.3–∞)

Table 3 – Paper I main association results. One-sided P-values and 95% confidence intervals were derived from Fisher's exact test.

Limitations of this study include the heterogeneous data sources, with different biochemical and computational methods employed for CNV calling. This is partly overcome by the application of a 100 kb lower length limit of CNVs included, since such large events are more consistently detected across platforms and algorithms. Furthermore, the consistency in size and inheritance of the CNVs in the discovery pedigree minimizes the risk of the initially detected variant being a false positive finding.

6.2 The *HOXB13* G84E mutation significantly increases PC risk

In paper II, we studied a recently discovered, coding, rare single nucleotide variant in the gene *HOXB13* associated with a highly increased risk of PC (reported OR ~20) [58]. Because the initial finding needed to be replicated in an independent sample, and because the initial estimate of effect size was very uncertain (95% CI 3.5–803, Fisher's exact test) due to very few control samples carrying the mutation, we embarked on a replication study using two Swedish case-control samples of PC: CAPS and Stockholm-1, described in sections 4.3 and 4.4.

The coding G84E variant (SNP ID rs138213197) and flanking common variants were successfully genotyped in 2,805 cases and 1,709 population controls in CAPS, and in 2,098 biopsy positive cases and 2,880 biopsy negative controls in Stockholm-1.

The G84E variant was found in higher frequencies in both Swedish samples studied, compared to the US-based samples initially analyzed by Ewing and colleagues. The carrier frequency in the Swedish control samples was similar to that in the US cases (1.3–1.4%), and even higher in the Swedish cases. The carrier frequencies found in CAPS and Stockholm-1 are displayed in table 4 along with the main association results.

The higher overall carrier frequency in the Swedish materials allowed us to infer more precise estimates of the relative risk increase associated with the G84E mutation in Sweden, and gave very similar estimates of 3.4 and 3.5 in CAPS and Stockholm-1 respectively. This places the *HOXB13* G84E mutation in the hitherto sparsely populated region of uncommon, moderately penetrant variants on the map of PC-associated genetic variation [57].

Sample	G84E noncarriers	G84E carriers	Carrier frequency	OR	95% CI	P
CAPS						
Controls	1,685	24	1.4%			
Cases	2,675	130	4.6%	3.4	(2.2–5.4)	6.4×10^{-10}
Stockholm-1						
Controls	2,843	37	1.3%			
Cases	2,007	91	4.3%	3.5	(2.4–5.2)	2.0×10^{-11}

Table 4 – Paper II main association results. Odds ratios, P-values and 95% confidence intervals were calculated by logistic regression.

Age	G84E noncarriers		G84E carriers		G84E carriers, risk score in Q4	
60	1.3%	(1.1–1.4)	3.9%	(2.5–6.2)	6.4%	(4.1–10.0)
65	3.3%	(3.0–3.6)	9.9%	(6.5–15.3)	16.1%	(10.6–24.4)
70	6.1%	(5.7–6.7)	18.1%	(12.1–27.0)	28.4%	(19.5–41.5)
75	9.2%	(8.6–9.9)	26.0%	(17.9–37.7)	39.6%	(28.4–55.3)
80	12.0%	(11.3–12.8)	32.5%	(23.1–45.9)	47.7%	(35.6–64.0)

Table 5 – Paper II absolute risk estimates up to certain ages with 95% confidence intervals. Estimates are adjusted for other cause mortality. Risks were estimated for G84E noncarriers, G84E carriers, and G84E carriers within the uppermost quartile of a genetic risk score based on 33 established PC risk SNPs.

Using register-based total population data on age-specific mortality and PC incidence, and the G84E mutation effect estimates from the population-based CAPS sample, we employed competing risks methods to estimate the lifetime absolute risk of PC in G84E carriers and noncarriers, adjusted for the competing risk of other cause mortality.

The lifetime absolute risk estimates up to certain ages for G84E carriers and noncarriers are presented in table 5. According to our model, almost a third of G84E carriers will be diagnosed with PC by age 80. In addition to the PC risk increase conferred by the HOXB13 G84E variant, we analyzed a model including G84E and quartiles of a genetic risk score based, on 33 established PC risk SNPs. Table 5 also shows the absolute risk estimates for G84E carriers within the uppermost quartile of this score, a combination expected to be found in ~0.3% of the Swedish population. In this group, the estimated lifetime PC risk up to age 80 reached almost 50%. In a hypothetical screening program where genetic markers were assessed, this group of men could be easily identified and offered more frequent subsequent screening or even genetic counseling.

6.3 Six genes are associated with TGCT risk, one modified by parent-of-origin

In paper III, we performed a combined case-parent, case-control association analysis of common SNPs in genes previously shown to be associated with TGCT. We furthermore investigated whether the estimated effects of SNPs were modified by parental sex, and tumor histological subtype (seminoma or non-seminoma).

118 SNP markers in or near the six genes *ATF7IP*, *BAK1*, *DMRT1*, *KITLG*, *SPRY4*, and *TERT* were genotyped in 831 case-parent triads, 474 case-parent dyads, and 712 singleton cases. Genome-wide SNP data, including the regions under investigation, were available for 3,919 unrelated male controls from the TwinGene project. Using imputation methods, genotypes were generated for the entire sample for a total of 852 SNP markers passing all quality control steps, including markers which were directly genotyped.

In a combined case-parent and case-control test for SNP-TGCT association, marker genotypes in all the investigated genes were found to be significantly associated with TGCT. The associations remained highly significant after false discovery rate (FDR) adjustment of P-values for multiple testing.

By gene-wise stepwise logistic regression among genotyped markers, 11 SNPs in total were found to be associated with TGCT when adjusting for the previously selected significant markers in the same gene. Three each of these were in the *TERT* and *BAK1* regions, two in the *SPRY4* region, and one each in *DMRT1*, *ATF7IP*, and *KITLG*. Figure 10 shows the overall association results and markers selected by stepwise regression per gene.

The 11 independently associated SNPs were tested for interaction of the allelic effect with parental sex, and with the tumor histological subtype (seminoma or non-seminoma) found in the proband. No markers showed significant statistical interaction with histological subtype, but one marker in the *SPRY4* region had a significantly different effect on disease risk depending on the sex of the parent from which an allele was inherited.

Specifically, the marker rs10463352, just upstream of the coding sequence for *SPRY4*, had an estimated OR of 1.72 (95% CI 1.38–2.15) for maternally inherited G alleles, while the same estimate for paternally inherited alleles was 0.99 (95% CI 0.70–1.39) (unadjusted P-value for interaction 0.0013). Thus, the overall allelic effect estimated for this marker was attenuated by including the paternally inherited alleles.

If the parent-of-origin effect seen for rs10463352 in TGCT holds up for replication in independent materials, this effect indicates gene silencing by genomic imprinting in males, or some other mechanism of gene-sex interaction. This is a new aspect of TGCT genetic epidemiology, and could lead to further insight in disease mechanisms.

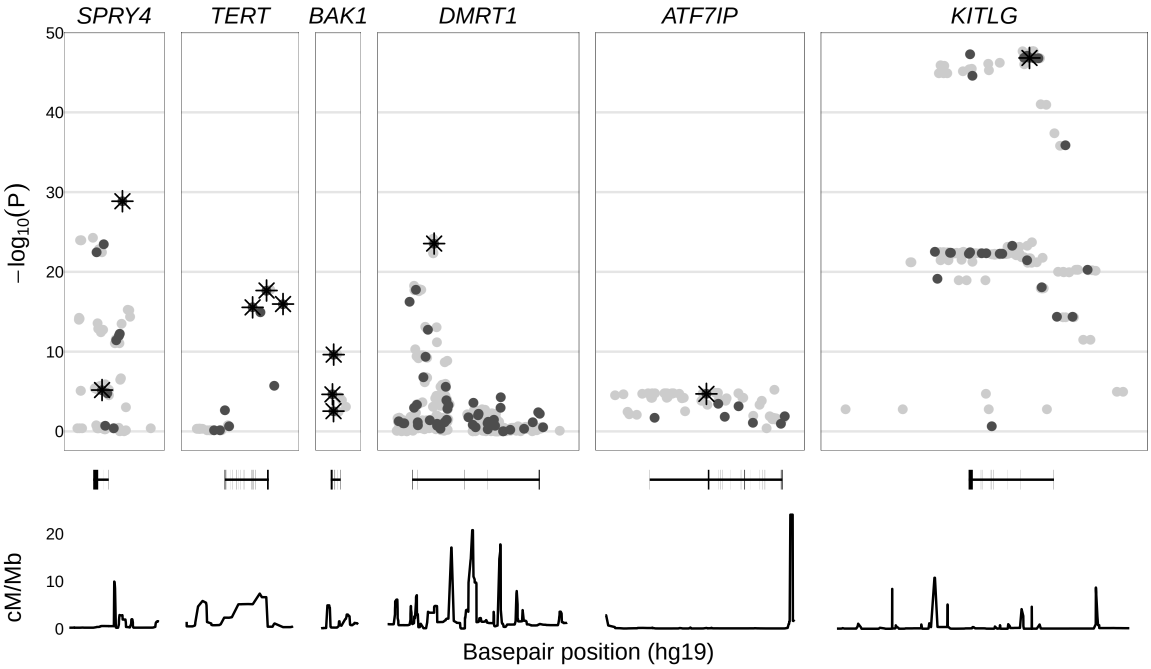

Figure 10 – Main association results of paper III. Black markers were directly genotyped, while gray markers were imputed. Star-shaped markers were selected in gene-wise forward stepwise regression. The middle panel shows the extents of transcribed regions of genes (horizontal lines) and exons (vertical lines). The lower panel shows recombination rate estimates from the 1000 genomes reference panel in centiMorgan/mega base pair units.

6.4 Untargeted UPLC-MS metabolomic analysis of serum could not reliably separate PC patients from controls

In paper IV, we performed a “metabolome-wide association study” in PC. The GWAS approach of screening an entire class of similar variants for association with disease or other traits was applied to the serum metabolome of PC patients and population controls.

Serum samples from 188 control subjects, 188 PC cases with indolent disease, and 99 PC cases with aggressive disease were selected from the CAPS biobank, and analyzed by UPLC-MS by collaborators at Colorado State University, Fort Collins, CO, USA. The UPLC-MS data were then postprocessed and transformed into quantitative measurements of 6,138 metabolite “features”.

Each feature corresponds to an intensity peak in the UPLC-MS chromatograms, which in turn correspond to a detected ion or ion fragment from the analyzed sample, or in some cases to noise. A single molecule species in the serum sample may correspond to several correlated peaks in the output due to fragmentation and the creation of adduct ions in the UPLC-MS process, and variations in isotope composition of the source molecule species.

In an exploratory analysis, intensities of many of the features were found to be associated with the age of the individual who had donated the serum sample. Furthermore, a significant portion of the feature intensities were associated with the length of time a serum sample had spent in frozen storage before analysis (range approximately 6–9 years). Since both these factors were also associated with disease status, they were adjusted for in the full regression analysis.

Feature–PC association was assessed using linear regression and the ANOVA F-test, testing whether each feature's intensity differed significantly between the three PC status categories, adjusted for sample storage time and age.

Two features were significantly associated with PC status after Bonferroni correction for multiple testing (P-values 4.0×10^{-6} and 7.1×10^{-6}). However, none of them could be identified as a specific molecule.

We further examined whether pairwise differences between feature intensities (corresponding to ratios between features as originally measured due to log-transformation of data) showed stronger association to disease status than single features. The rationale for this was that pairwise comparisons could potentially capture relations such as substrate–product in biochemical pathways. However, no such pairwise difference was significantly associated with PC status when considering the increased multiple testing burden of all pairwise comparisons. The strongest (but nonsignificantly associated) feature ratios contained molecules putatively identified as caprolactam, L-phosphatidic acid, and the tripeptide Tyr-Lys-Thr.

Finally, the four features showing the strongest association with PC status (not necessarily metabolome-wide significant) were tested for association with genotypes of ~1.4 million SNPs (genotyped and imputed) genome-wide. Traits were analyzed by linear regression of feature intensities on minor allele counts of genotypes, with no covariates. All four features had their most strongly associated markers in introns of genes. One feature (identifier 174.1_53) had a genome-wide significant association to a variant in the gene *IL13RA1*, which encodes a subunit of the interleukin 13 receptor, another subunit of which has been suggested as a drug target for prostate cancer treatment.

In conclusion, despite interesting auxiliary findings, the main aim of this study was to find novel biomarkers useful in assessing prostate cancer risk and aggressiveness. This aim was not fulfilled. The search performed within the scope of this paper was however far from exhaustive, and additional molecular detection strategies and samples may yet find PC biomarkers in the serum metabolome.

7 Conclusions and future perspective

7.1 Implications and future research based on the papers

Paper I Our multistage study provides strong evidence for association of rare *MAGI1* CNVs with disease, and the association seems even more plausible given the many interaction partners to MAGI1 that have been implied in psychiatric genetic epidemiology, and its role in the synapse (detailed in paper I with supplemental information). However, the rarity of these specific mutations may limit the utility of this knowledge. In future studies, common variation in the region could be reexamined, with the prior suspicion of involvement with BPAD and SZ lowering the required significance threshold compared to that required for a "hypothesis-free" GWAS.

Should the common variant approach not yield positive results, an alternative would be to perform deep resequencing of the *MAGI1* region in pedigrees with a high BPAD and/or SZ load, or even in unrelated cases and control subjects. Using the "next-generation" high-throughput sequencing technologies which have become available and affordable over the past few years, large numbers of subjects could be sequenced affordably. If *MAGI1* is important in BPAD and SZ pathways, there should be other rare variants than large CNVs affecting the gene's function in large cohorts of cases. A large scale resequencing study would clarify this.

A final intriguing lane of research would be to try and locate relatives of unrelated *MAGI1* CNV carriers, in order to increase understanding of the penetrance of these variants. If *MAGI1* CNVs are confirmed as highly penetrant susceptibility variants they would provide valuable markers for genetic counseling in families highly affected by BPAD and SZ.

Paper II We have shown that carriers of the HOXB13 G84E mutation are at considerable increased risk of developing PC, with one out of three carriers being affected during their lifetime. Our results agree with other replication studies in different populations, confirming the G84E mutation as strongly associated with prostate cancer [87, 88, 89, 90]. A fundamental difference in the Swedish population investigated herein is the relatively high prevalence of the mutation in control subjects (1.4%) as compared to other populations. In general, the mutation has been observed to be very rare in control populations outside the Nordic countries, supporting the hypothesis that G84E is a founder mutation of Nordic origin. This is further supported by our haplotype analysis, in which a single unique 108 kb haplotype carried the G84E risk allele in the Swedish population. Finally, a Chinese case-control study of *HOXB13* variants in PC found no carriers of G84E [91], and the International Consortium for

Prostate Cancer Genetics noted that none of their investigated families of African, Ashkenazi Jewish, or "Other" descent (excluding European) had any carriers of the variant [89].

To further explore the risk predictive capacity of G84E we stratified mutation carriers by a polygenic risk score composed of thirty-three established low-penetrant susceptibility variants. Mutation carriers in the top quartile of the risk score were at considerably elevated risk of PC with an estimated lifetime risk of 48%. Thus, integrating established PC susceptibility variants in risk prediction models may make targeted screening and intervention programs feasible, and additional efforts to translate PC genetic risk prediction into the clinical setting are highly warranted.

The identification of the *HOXB13* G84E mutation suggests the existence of other rare, moderately penetrant, susceptibility variants in PC genetics. Therefore, additional studies designed to identify such rare variants are highly warranted. Next-generation sequencing assessment of linked genomic regions for hereditary PC cases may provide an effective approach to identify additional rare variants.

Looking back to paper I, the findings of both studies indicate that new technologies such as CNV detection and deep resequencing applied to linkage regions from previously collected family materials can lead to new discoveries. The former study exemplifies the approach using CNVs and family-wise linkage, and the *HOXB13* G84E variant studied in paper II was first discovered by resequencing of genes in the 17q21-22 linkage region in PC families [58].

Paper III We have provided independent replication of previous GWAS findings for TGCT in a large, Scandinavian, case-parent, case-control sample. In addition to the straightforward replication, we also found indications of the effect of variants in the gene *SPRY4* on TGCT risk being modified by parental sex – the identity of the transmitted allele only made a difference for disease risk when inherited from mothers of cases. The mechanism behind this phenomenon is not clear. An interesting future study would be to assess genomic inactivation by methylation in the region where the parent-of-origin effect was detected, in cases as well as in their parents.

As for any other complex disease, there are most likely more risk loci of smaller effect waiting to be uncovered for TGCT. Considering that TGCT is the least common of the diseases investigated in this thesis, international collaborations may be required to find these variants. As a first step towards participation in an international meta-analysis, the cases and case-parent constellations analyzed in paper III should be genotyped on a genome-wide SNP chip, and a GWAS be performed.

Finally, the SNPs found to be associated with TGCT in paper III and in the GWASs motivating the study may not be the actual variants that cause the increase in disease risk, but only correlate with causal variants through LD. By resequencing the association regions in a subset of cases and controls, we may find variants of even stronger effect, which are actually causative of TGCT.

Paper IV We examined the UPLC-MS-detectable human serum metabolome for molecules associated with PC status, with the primary aim of finding molecules useful as biomarkers for PC aggressiveness. We were also interested in searching for any molecules that differed significantly in abundance in serum from PC cases compared to controls, even if not immediately useful as biomarkers, in order to inform us about disease mechanisms.

Two metabolomic features were significantly associated with disease status, but none of them could be unambiguously identified.

Future analyses could include analyzing the same samples by gas chromatography–mass spectrometry (GC–MS), which partly detects another spectrum of molecules, and with properties of the detection procedure making molecular identification of associated traits easier.

One could also imagine multivariate data mining or machine learning methods having better success in finding patterns separating cases from controls, than the simple regression models applied in paper IV. However, with more complex models come difficulties in interpreting their parameters, and in achieving independent replication.

The metabolome is a dynamic entity, constantly responding to environmental influences [92]. Compared to the relative stability of genomic DNA, there is thus much time-varying noise inherent in metabolomic analyses, which may hinder inference.

7.2 Future perspective on molecular/genetic epidemiology research

Assuming we knew the full genomic DNA sequence of everyone, which research questions could we answer? The somewhat unfortunately named "next generation" sequencing technologies (what should we call the next next generation?) have in just a few years delivered vast amounts of individual sequence data. In 2007 and 2008 the first personal genomes of James Watson [93], Craig Venter [94], and an anonymous Han Chinese donor [95] were presented as major scientific milestones, and the 1,000 genomes project (1kG) was initiated and announced its intent to sequence and make available for research at least one thousand human genomes [96]. Today, only four years later, the 1kG project has delivered the promised dense map of human genetic variation [97], and many personal genomes have been sequenced in clinical and research settings. Although the cost for sequencing a personal genome has not yet plummeted to the 1,000 (US) dollar genome, it seems to be on its way there.

In a clinical personal genomic era we will undoubtedly find unexpected disease-associated variation in almost everyone. If whole genome sequencing is to become routine as part of a screening program, we need a much greater understanding of which variants are actually harmful (e.g. BRCA1 mutations in breast cancer), and which are not (e.g. any of a large number of theoretically deleterious but in practice neutral variants discovered through whole genome sequencing [98]). Furthermore, if no action can be taken on a discovered allele increasing risk for a disease, would the (prospective) patient want to know that he or she carries it? In these still early days, genomic sequence will probably be of greater use in the research laboratory than in the clinic. There, molecular mechanisms of disease can be elucidated, and treatment strategies based on these mechanisms can be devised.

8 Acknowledgements

Many people have been involved and invaluable in my journey from registration to dissertation, and I hope I have not forgotten too many of you. My sincerest thanks to:

Fredrik Wiklund, my current main supervisor, who with incurable optimism and Norrlandish realism has guided me through the second half of my thesis work.

Silvia Paddock, my former main supervisor, who picked me up from the mathematics department, taught me all about human genetics, and got me on the Ph.D. train.

Juni Palmgren and Jonathan Prince, my co-supervisors. Juni impressed me from day one, and through her immense connectedness and academic matchmaking skills I ended up at the department of Medical Epidemiology and Biostatistics for the second half of my thesis work. Jonathan, though no longer at the department, keeps inspiring from a distance.

Magnus Svedjebratt, my external mentor, always ready to provide good advice on request.

Fellow, former, and future friendly graduate students, biostatisticians, staff and faculty at MEB, including but not limited to: Robert Szulkin, Caroline Weibull, Henrik Olsson, Emil Rehnberg, Sandra Eloranta, Therese Andersson, Anna Johansson, Hatef Darabi, Mun-Gwan Hong, Linda Abrahamsson, Iffat Rahman, Ralf Kuja-Halkola, Thomas Frisell, Lovisa Högberg, Adina Feldman, Martin Fransson, Stephanie Bonn, Sara Christensen, Lisa Möller, Maria Sandberg, Andrea Ganna, Johanna Holm, Karin Sundström, Amy Levál, Miriam Elfström, Vilhelmina Ullemar, Ci Song, Markus Aly, Therese Ljung, Mårten Neiman, Martin Eklund, Anna Kähler, Jitender Kumar, Mark Clements, Keith Humphreys, Yudi Pawitan, Marie Reilly, Marie Jansson, Gunilla Sonnebring, Erika Nordenhagen, Sven Sandin, Alex Ploner, Arvid Sjölander, Cecilia Lundholm, Paul Dickman, Patrik Magnusson, Kamila Czene. Thank you for good company and inspiration along the way!

Henrik Grönberg, head of the department of Medical Epidemiology and Biostatistics.

Friends and faculty from the department of Neuroscience: Lisette Graae, Magnus Lekman, Caroline Ran, Anna Anvret, Mimi Westerlund, Andrea Carmine Belin, Dagmar Galter, Mathew Abrams, Susanne Szydlowski, Sophia Savage, Lars Olson, Karin Pernold, Karin Lundströmer, Eva Lindqvist. Thank YOU for good company along the way!

Ida Engqvist, IT coordinator at the department of Neuroscience, and a source of endless IT wisdom and wizardry.

The KIRC group: Helga Westerlind, Ryan Ramanujam, Henrik Källberg, Johan Acevedo, Boel Brynedal, Daniel Uvehag, Jonas Forsslund. Keep KI nerdy!

Ulf Rydmark who taught me mathematics, och Per Cederholm som lärde mig ord.

The MEB IT group, who always keep the servers humming. Special thanks to resident Linux gurus Zack Yusuf and Rikard Öberg.

Collaborators in Norway: Tom Grotmol, Trine Haugen, Wenche Kristiansen, Kristine Andreassen, Elin Aschim.

Family: Ulf Karlsson, Lena Malmstedt, Ellinor Larsson, Per Ekholm. Gärd Allvin who always kept me well fed, Karin and Yngve Allvin who always managed to seem interested in what I've been doing.

Helen ♡ ∞

References

[1] Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171(4356):737–738.

[2] Watson JD, Crick FHC. Genetical Implications of the Structure of Deoxyribonucleic Acid. Nature. 1953;171(4361):964–967.

[3] Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561–563.

[4] Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321(6067):209–213.

[5] Hannon GJ. RNA interference. Nature. 2002;418(6894):244–251.

[6] Nagano T, Fraser P. No-Nonsense Functions for Long Noncoding RNAs. Cell. 2011;145(2):178–181.

[7] Huarte M, Rinn JL. Large non-coding RNAs: missing links in cancer? Human Molecular Genetics. 2010;19(R2):R152–R161.

[8] Mendel GV. Versuche Über Pflanzen-Hybriden. Journal of Heredity. 1951;42(1):3–4.

[9] Wiener AS. Method of Measuring Linkage in Human Genetics; with Special Reference to Blood Groups. Genetics. 1932;17(3):335–350.

[10] Ott J. Analysis of Human Genetic Linkage. Johns Hopkins University Press; 1999.

[11] Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. Nature Genetics. 1997;17(1):21–24.

[12] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29(1):308–311.

[13] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science. 2004;305(5683):525–528.

[14] Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. Science. 1996;273(5281):1516–1517.

[15] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005;308(5720):385–389.

[16] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–678.

[17] Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. American Journal of Human Genetics. 2012;90(1):7–24.

[18] National Human Genome Research Institute. Talking Glossary of Genetic Terms. National Institutes of Health;. Available from: `http://www.genome.gov/glossary/?id=114`.

[19] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

[20] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001;291(5507):1304–1351.

[21] Rice TK, Borecki IB. Familial resemblance and heritability. In: Advances in Genetics. vol. Volume 42. Academic Press; 2001. p. 35–44.

[22] Maher B. Personal genomes: The case of the missing heritability. Nature News. 2008;456(7218):18–21.

[23] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753.

[24] Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. 2012;109(4):1193–1198.

[25] Hardy GH. Mendelian Proportions in a Mixed Population. Science. 1908;28(706):49–50.

[26] American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-IV. American Psychiatric Association; 1994.

[27] Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet. 2009;373(9659):234–239.

[28] The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748–752.

[29] Craddock N, Owen MJ. The Kraepelinian dichotomy - going, going... but still not gone. The British Journal of Psychiatry. 2010;196(2):92–95.

[30] McGrath J, Saha S, Chant D, Welham J. Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. Epidemiologic Reviews. 2008;30(1):67–76.

[31] Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Archives of general psychiatry. 2011;68(3):241–251.

[32] Osby U, Brandt L, Correia N, Ekbom A, Sparén P. Excess mortality in bipolar and unipolar disorder in Sweden. Archives of general psychiatry. 2001;58(9):844–850.

[33] McGrath JJ. Variations in the Incidence of Schizophrenia: Data Versus Dogma. Schizophrenia Bulletin. 2006;32(1):195–197.

[34] Tsuchiya KJ, Byrne M, Mortensen PB. Risk factors in relation to an emergence of bipolar disorder: a systematic review. Bipolar Disorders. 2003;5(4):231–242.

[35] Frans EM, Sandin S, Reichenberg A, Lichtenstein P, Långström N, Hultman CM. Advancing paternal age and bipolar disorder. Archives of general psychiatry. 2008;65(9):1034–1040.

[36] Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nature Reviews Genetics. 2012;13(8):537–551.

[37] Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CAM, et al. Disruption of two novel genes by a translocation co-segregating with schizophrenia. Human Molecular Genetics. 2000;9(9):1415–1423.

[38] Sanders M, Duan P, Levinson M, Shi P, He B, Hou B, et al. No Significant Association of 14 Candidate Genes With Schizophrenia in a Large European Ancestry Sample: Implications for Psychiatric Genetics. American Journal of Psychiatry. 2008;165(4):497–506.

[39] Kim Y, Zerwas S, Trace SE, Sullivan PF. Schizophrenia Genetics: Where Next? Schizophrenia Bulletin. 2011;37(3):456–463.

[40] The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008;455(7210):237–241.

[41] The Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43(10):969–976.

[42] Sklar P, Smoller J, Fan J, Ferreira M, Perlis R, Chambert K, et al. Whole-genome association study of bipolar disorder. Molecular Psychiatry. 2008;13(6):558–569.

[43] Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics. 2011;43(10):977–983.

[44] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: A Cancer Journal for Clinicians. 2011;61(2):69–90.

[45] Engholm G, Ferlay J, Christensen N, Bray F, Gjerstorff ML, Klint Å, et al. NORDCAN – a Nordic tool for cancer information, planning, quality control and research. Acta Oncologica. 2010;49(5):725–736.

[46] Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-Specific Antigen as a Serum Marker for Adenocarcinoma of the Prostate. New England Journal of Medicine. 1987;317(15):909–916.

[47] Welch HG, Black WC. Overdiagnosis in Cancer. Journal of the National Cancer Institute. 2010;102(9):605–613.

[48] Nordström T, Aly M, Clements MS, Weibull CE, Adolfsson J, Grönberg H. Prostate-specific Antigen (PSA) Testing Is Prevalent and Increasing in Stockholm County, Sweden, Despite No Recommendations for PSA Screening: Results from a Population-based Study, 2003–2011. European Urology. 2012;[Epub ahead of print]:http://dx.doi.org/10.1016/j.eururo.2012.10.001.

[49] Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Ciatto S, Nelen V, et al. Prostate-cancer mortality at 11 years of follow-up. The New England journal of medicine. 2012;366(11):981–990.

[50] Andriole GL, Crawford ED, Grubb RL, Buys SS, Chia D, Church TR, et al. Prostate Cancer Screening in the Randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: Mortality Results after 13 Years of Follow-up. Journal of the National Cancer Institute. 2012;104(2):125–132.

[51] Grönberg H. Prostate cancer epidemiology. The Lancet. 2003;361(9360):859–864.

[52] Venkateswaran V, Klotz LH. Diet and prostate cancer: mechanisms of action and implications for chemoprevention. Nature Reviews Urology. 2010;7(8):442–453.

[53] Hedelin M, Chang ET, Wiklund F, Bellocco R, Klint Å, Adolfsson J, et al. Association of frequent consumption of fatty fish with prostate cancer risk is modified by COX-2 polymorphism. International Journal of Cancer. 2007;120(2):398–405.

[54] Hedelin M, Klint Å, Chang E, Bellocco R, Johansson JE, Andersson SO, et al. Dietary Phytoestrogen, Serum Enterolactone and Risk of Prostate Cancer: The Cancer Prostate Sweden Study (Sweden). Cancer Causes and Control. 2006;17(2):169–180.

[55] Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. New England Journal of Medicine. 2000;343(2):78–85.

[56] Sundararajan S, Ahmed A, Goodman J Oscar B. The relevance of BRCA genetics to prostate cancer pathogenesis and treatment. Clinical Advances in Hematology & Oncology: H&O. 2011;9(10):748–755.

[57] Goh CL, Schumacher FR, Easton D, Muir K, Henderson B, Kote-Jarai Z, et al. Genetic variants associated with predisposition to prostate cancer and potential clinical implications. Journal of Internal Medicine. 2012;271(4):353–365.

[58] Ewing CM, Ray AM, Lange EM, Zuhlke KA, Robbins CM, Tembe WD, et al. Germline Mutations in HOXB13 and Prostate-Cancer Risk. New England Journal of Medicine. 2012;366(2):141–149.

[59] Bosl GJ, Motzer RJ. Testicular Germ-Cell Cancer. New England Journal of Medicine. 1997;337(4):242–254.

[60] Rosen A, Jayram G, Drazer M, Eggener SE. Global Trends in Testicular Cancer Incidence and Mortality. European Urology. 2011;60(2):374–379.

[61] Hemminki K, Li X. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. British Journal of Cancer. 2004;90(9):1765–1770.

[62] Peng X, Zeng X, Peng S, Deng D, Zhang J. The Association Risk of Male Subfertility and Testicular Cancer: A Systematic Review. PLoS ONE. 2009;4(5):e5591.

[63] Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. International Journal of Cancer. 2002;99(2):260–266.

[64] Nathanson KL, Kanetsky PA, Hawes R, Vaughn DJ, Letrero R, Tucker K, et al. The Y Deletion gr/gr and Susceptibility to Testicular Germ Cell Tumor. The American Journal of Human Genetics. 2005;77(6):1034–1043.

[65] Rapley EA, Turnbull C, Al Olama AA, Dermitzakis ET, Linger R, Huddart RA, et al. A genome-wide association study of testicular germ cell tumor. Nature Genetics. 2009;41(7):807–810.

[66] Kanetsky PA, Mitra N, Vardhanabhuti S, Vaughn DJ, Li M, Ciosek SL, et al. A Second Independent Locus Within DMRT1 Is Associated with Testicular Germ Cell Tumor Susceptibility. Human Molecular Genetics. 2011;20(15):3109–3117.

[67] Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, et al. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. Nature Genetics. 2009;41(7):811–815.

[68] Turnbull C, Rapley EA, Seal S, Pernet D, Renwick A, Hughes D, et al. Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. Nature Genetics. 2010;42(7):604–607.

[69] Chung C, Wang Z, Kanetsky P, Turnbull C, McGlynn K, Erickson R, et al. Meta-analysis identifies four new loci for testicular germ cell tumor. In: the 62nd Annual Meeting of The American Society of Human Genetics. San Francisco, CA, USA; 2012. p. #213.

[70] The National Institute of Mental Health. NIMH Bipolar Disorder Genetics Initiative; [Accessed 13th December 2012]. [Online]. Available from: `http://www.nimhgenetics.org/available_data/bipolar_disorder/`.

[71] Olama AAA, Kote-Jarai Z, Schumacher FR, Wiklund F, Berndt SI, Benlloch S, et al. A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. Human Molecular Genetics. 2012;[Epub ahead of print]:http://dx.doi.org/10.1093/hmg/dds425.

[72] Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Human Molecular Genetics. 2011;20(19):3867–3875.

[73] Aly M, Wiklund F, Xu J, Isaacs WB, Eklund M, D'Amato M, et al. Polygenic Risk Score Improves Prostate Cancer Risk Prediction: Results from the Stockholm-1 Cohort Study. European Urology. 2011;60(1):21–28.

[74] Kristiansen W, Andreassen KE, Karlsson R, Aschim EL, Bremnes RM, Dahl O, et al. Gene Variations in Sex Hormone Pathways and the Risk of Testicular Germ Cell Tumour: A Case–Parent Triad Study in a Norwegian–Swedish Population. Human Reproduction. 2012;27(5):1525–1535.

[75] Rahman I, Bennet AM, Pedersen NL, de Faire U, Svensson P, Magnusson PKE. Genetic Dominance Influences Blood Biomarker Levels in a Sample of 12,000 Swedish Elderly Twins. Twin Research and Human Genetics. 2009;12(03):286–294.

[76] Ragoussis J. Genotyping Technologies for Genetic Research. Annual Review of Genomics and Human Genetics. 2009;10(1):117–133.

[77] Kong A, Cox NJ. Allele-Sharing Models: LOD Scores and Accurate Linkage Tests. The American Journal of Human Genetics. 1997;61(5):1179–1188.

[78] Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research. 2007;17(11):1665–1674.

[79] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature Protocols. 2010;5(9):1564–1573.

[80] Lambert CG, Black LJ. Learning from our GWAS mistakes: from experimental design to scientific method. Biostatistics. 2012;13(2):195–203.

[81] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81(3):559–75.

[82] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

[83] Browning BL, Browning SR. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics. 2009;84(2):210–223.

[84] Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics. 2007;81(5):1084–1097.

[85] Dudbridge F. Likelihood-Based Association Analysis for Nuclear Families and Unrelated Subjects with Missing Genotype Data. Human heredity. 2008;66(2):87–98.

[86] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. Analytical Chemistry. 2006;78(3):779–787.

[87] Akbari MR, Trachtenberg J, Lee J, Tam S, Bristow R, Loblaw A, et al. Association Between Germline HOXB13 G84E Mutation and Risk of Prostate Cancer. Journal of the National Cancer Institute. 2012;104(16):1260–1262.

[88] Breyer JP, Avritt TG, McReynolds KM, Dupont WD, Smith JR. Confirmation of the HOXB13 G84E Germline Mutation in Familial Prostate Cancer. Cancer Epidemiology Biomarkers & Prevention. 2012;21(8):1348–1353.

[89] Xu J, Lange E, Lu L, Zheng S, Wang Z, Thibodeau S, et al. HOXB13 is a susceptibility gene for prostate cancer: results from the International Consortium for Prostate Cancer Genetics (ICPCG). Human Genetics. 2012;[Epub ahead of print]:http://dx.doi.org/10.1007/s00439–012–1229–4.

[90] Stott-Miller M, Karyadi DM, Smith T, Kwon EM, Kolb S, Stanford JL, et al. HOXB13 mutations in a population-based, case–control study of prostate cancer. The Prostate. 2012;[Epub ahead of print]:http://dx.doi.org/10.1002/pros.22604.

[91] Lin X, Qu L, Chen Z, Xu C, Ye D, Shao Q, et al. A novel Germline mutation in HOXB13 is associated with prostate cancer risk in Chinese men. The Prostate. 2012;[Epub ahead of print]:http://dx.doi.org/10.1002/pros.22552.

[92] Krug S, Kastenmuller G, Stückler F, Rist MJ, Skurk T, Sailer M, et al. The dynamic range of the human metabolome revealed by challenges. The FASEB Journal. 2012;26(6):2607–2619.

[93] Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452(7189):872–876.

[94] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The Diploid Genome Sequence of an Individual Human. PLoS Biol. 2007;5(10):e254.

[95] Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. Nature. 2008;456(7218):60–65.

[96] The 1000 Genomes Project. International Consortium Announces the 1000 Genomes Project; 2008. Press Release. Available from: http://www.1000genomes.org/sites/1000genomes.org/files/docs/1000genomes-newsrelease.pdf.

[97] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.

[98] Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, et al. Deleterious- and Disease-Allele Prevalence in Healthy Individuals: Insights from Current Predictions, Mutation Databases, and Population-Scale Resequencing. The American Journal of Human Genetics. 2012;91(6):1022–1032.