

## **Integrative data analysis of male reproductive disorders**

**Edsgard, Stefan Daniel; Brunak, Søren; Jensen, Thomas Skøt**

*Publication date:*  
2012

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

*Citation (APA):*  
Edsgard, S. D., Brunak, S., & Jensen, T. S. (2012). Integrative data analysis of male reproductive disorders. Kgs. Lyngby: Technical University of Denmark (DTU).

## **DTU Library** Technical Information Center of Denmark

---

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# **Integrative data analysis of male reproductive disorders**

– PhD Thesis –

**Daniel Edsgård**

Center for Biological Sequence Analysis  
Department of Systems Biology  
Technical University of Denmark

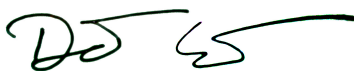
March 5, 2011



## Preface

This thesis was prepared at the Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark. It was carried out under the supervision of Professor Søren Brunak and Associate Professor Thomas Skøt Jensen. The work was made possible by a grant from the Villum Kann Rasmussen foundation.

Daniel Edsgård  
Lyngby, February 2011

A handwritten signature in black ink, consisting of stylized letters 'D', 'E', and 'S' followed by a long horizontal flourish.

---

# Contents

---

Preface . . . . .	iii
Contents . . . . .	iv
Abstract . . . . .	vii
Svensk resumé . . . . .	viii
Acknowledgements . . . . .	ix
Papers included in the thesis . . . . .	x
Papers not included in the thesis . . . . .	x
<b>I Introduction</b>	<b>1</b>
<b>1 Prelude</b>	<b>3</b>
1.1 Motivation and context . . . . .	3
1.2 Organization . . . . .	4
<b>2 Complex phenotypes and genetic variation</b>	<b>5</b>
2.1 Disease genetics . . . . .	5
2.2 Human genetic variation . . . . .	8
2.3 The cancer genome . . . . .	9
2.4 Testicular dysgenesis syndrome . . . . .	9
<b>3 Molecular characterization of complex phenotypes</b>	<b>15</b>
3.1 Genome-wide association studies . . . . .	15
3.2 Transcriptomics . . . . .	17
3.3 Protein interactomics . . . . .	18
3.4 Chemical biology . . . . .	19
3.5 Integrative systems biology . . . . .	21
<b>4 Technical and analytical aspects</b>	<b>23</b>
4.1 Microarrays . . . . .	23
4.2 Protein interactomics . . . . .	29
4.3 Digital gene expression . . . . .	30
4.4 Analysis . . . . .	31

<b>II Papers</b>	<b>35</b>
<b>5 Expression profiling of testicular carcinoma <i>in situ</i></b>	<b>37</b>
5.1 Abstract . . . . .	37
5.2 Manuscript . . . . .	38
5.3 Supplementary Material . . . . .	52
<b>6 Familial copy number variation in testicular cancer</b>	<b>55</b>
6.1 Abstract . . . . .	55
6.2 Manuscript . . . . .	56
6.3 Supplementary Material . . . . .	69
<b>7 Genome-wide association study of men with testicular dysgenesis syndrome</b>	<b>71</b>
7.1 Abstract . . . . .	71
7.2 Manuscript . . . . .	72
7.3 Supplementary material . . . . .	86
<b>8 Rare copy number variations affecting the risk for testicular germ cell tumor</b>	<b>87</b>
8.1 Abstract . . . . .	87
8.2 Manuscript . . . . .	88
8.3 Supplementary material . . . . .	97
<b>9 Next generation sequencing: RNA-seq of an unsequenced plant</b>	<b>99</b>
9.1 Abstract . . . . .	99
9.2 Manuscript . . . . .	100
9.3 Supplementary material . . . . .	121
<b>III Epilogue</b>	<b>123</b>
<b>10 Concluding remarks</b>	<b>125</b>
10.1 Summary and perspectives . . . . .	125
10.2 Outlook . . . . .	127
<b>Bibliography for Chapter 1</b>	<b>129</b>
<b>Bibliography for Chapter 2</b>	<b>130</b>
<b>Bibliography for Chapter 3</b>	<b>133</b>
<b>Bibliography for Chapter 4</b>	<b>136</b>
<b>Bibliography for Chapter 5</b>	<b>139</b>
<b>Bibliography for Chapter 6</b>	<b>142</b>

<b>Bibliography for Chapter 7</b>	<b>145</b>
<b>Bibliography for Chapter 8</b>	<b>148</b>
<b>Bibliography for Chapter 9</b>	<b>151</b>
<b>Bibliography for Chapter 10</b>	<b>157</b>
<b>Appendix A. Supplementary material paper II</b>	<b>158</b>
<b>Appendix B. Supplementary material paper III</b>	<b>163</b>
<b>Appendix C. Supplementary material paper IV</b>	<b>175</b>

## Abstract

During the last decades a decline in male reproductive health has been observed in Nordic countries, and particularly in Denmark. Testicular cancer is the most fatal form of male reproductive disorders, and despite high remission rates it is typically accompanied with infertility. The main topic of this thesis is the identification of the molecular basis of male reproductive disorders, with a special focus on testicular cancer. To this end, clinical samples were characterized by microarray-based transcription and genomic variation assays and molecular entities were identified by computational analysis of such data in conjunction with data from publicly available repositories. This thesis presents an introduction to disease genetics and molecular systems biology, followed by four studies that each provide detailed clues to the etiology of male reproductive disorders. Finally, a fifth study illustrates the use of massively parallel nucleotide sequencing for gene expression analysis.

In paper I the similarity of testicular carcinoma *in situ* cells to cells from the developing testis was investigated. We observed a close similarity to gonocytes, contributing further indications that non-spermatocytic testicular cancer arise due to disturbances in early testicular development. In paper II we analysed copy number variations (CNV) in germline DNA from four families with testicular germ cell tumors. Given the low number of samples, we aimed to improve the confidence by placing CNVs in a protein network context superimposed with established phenomic information. We thereby identified a recurrent CNV at a locus with genes encoding for the relaxin peptide hormones, indicating their potential role in testis function. Paper III presents a genome-wide association study on testicular dysgenesis syndrome. We confirmed the importance of *KITLG* in testicular cancer, and identified two risk loci related to the TGF $\beta$ -signaling pathway, *TGFBR3* and *BMP7*, by using a systems biology approach that was guided by the developmental disease hypothesis, and a pathway analysis based approach, respectively. Paper IV investigate the genome-wide association data with respect to copy number variation and show that the aggregated effect of rare variants can influence the risk for testicular cancer. Paper V provides an example of the application of RNA-Seq for expression analysis of a species with an unsequenced genome. We analysed the plant *Craterostigma plantagineum*, which is known for its astonishing drought tolerance, and thereby provided the first transcriptomes of this species. Comparisons of unstressed to desiccated conditions indicated several pathways of interest.

In conclusion, this thesis contributes to the molecular understanding of testicular malfunction and desiccation tolerance in *C. plantagineum*, as well as develops and highlights the usefulness of novel systems biology methodologies.



## Svensk resumé

De senaste decennierna har uppvisat en negativ trend vad gäller reproduktiv hälsa hos män i Norden, och Danmark är särskilt drabbat. Testikelcancer är den mest fatale formen bland manliga reproduktionssjukdomar, och trots en hög grad av remission är den ofta medföljd av infertilitet. Huvudämnet för denna doktorsavhandling är identifieringen av de molekylära faktorer som ligger bakom reproduktionsrelaterade sjukdomar hos män. För detta ändamål karakteriserades kliniska prover med avseende på genetisk transkription och genetisk variation genom mikroarraybaserade teknologier, och molekylära faktorer identifierades med hjälp av avancerad datoranalys av sådan data, samt samman med publikt tillgänglig data. Denna avhandling framlägger en introduktion till sjukdomsgenetik och molekylär systembiologi, följt av fyra studier som ger detaljerade ledtrådar till ovan nämnda sjukdomars etiologi. Slutligen presenteras en femte studie vilken illustrerar användandet av massiv parallelsekvensering för analys av genexpression.

Artikel I undersöker likheten mellan celler från carcinoma *in situ* och celler från testikel under embryonal utveckling. Nära likhet med gonocyter observerades, vilket bidrar med ytterligare indikationer på att icke-spermatocytisk testikelcancer uppstår på grund av störningar under tidig testikelutveckling. Artikel II analyserar kopietalsförändringar i nedärvt DNA från fyra familjer med testikeltumörer. Givet det låga antalet individer, sökte vi förbättra tillförlitligheten genom att integrera funna kopietalsvariationer i proteinnätverk överlagrat av fenotypisk information. En kopietalsvariation återfanns vid ett lokus kodande för peptidhormoner kallade relaxiner, vilket indikerar deras potentiella roll i testikelfunktion. Artikel III presenterar en genomvid associationsstudie. Vi bekräftade *KITLG*'s betydelse i testikelcancer, och fann två risk-associerade loci involverade i TGF $\beta$ -signalering, *TGFBR3* samt *BMP7*, genom tillämpning av en systembiologisk metod. I artikel IV undersöktes den genomvida datan med avseende på kopietalsvariation och det påvisas att den samlade påverkan av sällsynta varianter kan influera risken för testikelcancer. Artikel V exemplifierar användandet av massiv sekvenserings teknologi, genom genexpressionsanalys av en osekvenserad växt, *Craterostigma plantagineum*, känd för sin extrema tolerans mot dehydrering. Vi tillhandahåller de första hela transkriptomen för denna planta och pekar på flera signalvägar involverade i tolerans mot torka.

Sammantaget bidrar denna avhandling till förståelsen av testikulära funktionsstörningar och tolerans mot dehydrering i *C. plantagineum*, samt utvecklar och belyser nyttan av nya systembiologiska analysmetoder.

## Acknowledgements

I would like to express my gratitude to all those who have directly or indirectly contributed to this work. I would especially like to thank:

- My supervisor Søren Brunak for initiating this project and his enthusiasm, encouragement and great ideas. A huge thank you for creating the exciting and inspiring research environment that CBS is and for providing highly appreciated extra-curricular activities.
- My co-supervisor Thomas Skøt Jensen for continuously fruitful help and support along the way, and Henrik Bjørn Nielsen for excellent supervision on the plant project.
- A multitude of colleagues at CBS who I have been blissed to get to know, discuss with, and learn from, but it has been a unique pleasure to make friends with Nils, Thomas, Konrad and Tune, and to daily meet you.
- Our external collaborators at the Department of Growth and Reproduction, Rigshospitalet, Copenhagen, for their outstanding expertise in male reproductive health, critical assessments and experimental work and validation, as well our collaborators at the Department of Biostatistics, University of Copenhagen, for statistical help and implementations. My external collaborators within the plant project at the Department of Biology, University of Copenhagen, for high efficiency and scientific skill.
- The CBS administration for all help with formalities and the CBS system administration for maintaining the heavily burdened computer infrastructures.
- My friends and family, and especially my brother and sister, who revitalize me when I need it the most.

## Papers included in the thesis

- I Sonne SB, Almstrup K, Dalgaard MD, Juncker AS, **Edsgård D**, Ruban L, Harrison JN, Schwager C, Abdollahi A, Huber PE, Brunak S, Gjerdrum LM, Moore HD, Andrews PW, Skakkebæk NE, Rajpert-De Meyts E, and Leffers H. Analysis of gene expression profiles of microdissected cell populations indicates that testicular carcinoma *in situ* is an arrested gonocyte. *Cancer Research*, 2009
- II **Edsgård D**, Scheel M, Tue Hansen N, Skøt Jensen T, Gupta R, Brunak S, Skakkebæk NE, Rajpert-De Meyts E, and Ottesen AM. Heterozygous deletion at the RLN1 locus in a family with testicular germ cell cancer identified by integrating copy number variation data with phenome and interactome information. (Submitted)
- III Dalgaard MD<sup>†</sup>, Weinhold N<sup>†</sup>, **Edsgård D**<sup>†</sup>, Silver J, Pers TH, Jørgensen N, Juul A, Gerds TA, Giwercman A, Giwercman YL, Cedermark GC, Virtanen HE, Toppari J, Daugaard G, Skøt Jensen T, Brunak S, Rajpert-De Meyts E, Skakkebæk NE, Leffers H, and Gupta R. A genome-wide association study of men with testicular dysgenesis syndrome. (Submitted)
- IV **Edsgård D**, Dalgaard MD, Weinhold N, Jørgensen N, Juul A, Gupta R, Skøt Jensen T, Rajpert-De Meyts E, Leffers H, Skakkebæk NE, and Brunak S. Copy number variations affecting risk for testicular dysgenesis syndrome. (In preparation)
- V Suarez Rodriguez MC<sup>†</sup>, **Edsgård D**<sup>†</sup>, Hussain SS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D, and Mundy J. Transcriptomes of the desiccation tolerant resurrection plant *Craterostigma plantagineum*. *The Plant Journal*, 2010

(<sup>†</sup>) These authors contributed equally

## Papers not included in the thesis

- VI Taboureau O, Nielsen S, Audouze K, Weinhold W, **Edsgård D**, Roque F, Kouskoumvekaki I, Bora A, Curpan R, Skøt T, Brunak S and Oprea T. ChemProt: Integrating chemical-protein associations in phenome-interactome networks. *Nucleic Acids Research*, 2010

## **Part I**

# **Introduction**



---

# Chapter 1

## Prelude

---

### 1.1 Motivation and context

Cancer is responsible for one in eight deaths worldwide.<sup>1</sup> Genetics has been successful in identifying the causes behind rare disorders during the last decades, but common diseases, such as cancer, still present a challenging task.<sup>2</sup> Large international collaborative efforts have been undertaken that catalog genetic variation in populations, clinical cohorts and cancers. Concurrently, we are provided more and more complete maps of the proteome, transcriptome and epigenome. This poses a question, how can we make use of all available information?

The complexity of common diseases confronts us with a range of questions: What is the amount of clinical incidences that can be attributed to a particular genetic locus? How are genetic variations coupled to the endpoint phenotype, for example, what are the effects on a transcriptional and signaling level? What is the joint effect of a set of variations, and the impact of different genetic backgrounds? How do we handle the enormous search space when the measured number of molecular features are order of magnitudes larger than the number of available samples? What criteria should be used to define a phenotype when the underlying trait in reality is continuous?

These questions illustrate that common diseases constitute a more complex etiology than rare monogenic diseases. It also suggests that a more integrated view may be beneficial where the interplay and regulation between several biological entities of a cellular system is taken into account. In line with this, the aim of this thesis was to analyse clinical large-scale datasets, and combine it with available catalogs of information as to highlight molecular systems of interest, by the use of computational systems biology.

The work was performed in close collaboration with the Department of Growth and Reproduction, Rigshospitalet, Copenhagen, who generated the data and handled the cohorts. Even though this thesis involves basic research, the closeness to

clinical expertise assures that the findings are translated into clinical settings whenever possible, and nurtures the hope that the research can be put into practice for the good of patients and future prevention.

## 1.2 Organization

Part I gives an introduction to disease genetics and molecular systems biology. This chapter provides a birds-view of this thesis and its organization. Chapter 2 provides a background to genetics, the recent state of knowledge of human genetic variation, and a brief background to the testicular disorders that were studied in this thesis. Chapter 3 describes properties of the omics data that was used in this thesis, including data from the genome, transcriptome and proteome, as well as highlighting the importance of chemical influences from the environment. Concluding the chapter, the rationale of systems biology and integration of omics data is introduced. Chapter 4 describes experimental platforms and a background to the data analysis performed in part II.

Part II describes four studies that each provide detailed clues to the etiology of male reproductive disorders. Finally, a fifth study illustrates the use of massively parallel nucleotide sequencing for gene expression analysis.

Part III contains summarizing perspectives and states a few possible future directions for each of the studies presented. Finally, a few current issues in and general directions of human medical genetics research are outlined.

---

## Chapter 2

# Complex phenotypes and genetic variation

---

This chapter provides a background to the genetics of complex phenotypes, the recent state of knowledge of human genetic variation, and an introduction to the testicular disorders that were studied in this thesis.

### 2.1 Disease genetics

Genetic disorders are caused by effects that can be traced to the genetic makeup of an individual. Typically, these diseases can be inherited by passing on the genetic factors that influence susceptibility for disease from one generation to the next. Genetic disorders can be classified as either rare monogenic *Mendelian disorders* or otherwise *complex diseases*, depending on their genetic architecture and inheritance. The background to this distinction is presented in the following subsections.

#### Mendel and classical genetics

The Mendelian laws of inheritance are statements describing how heritable traits are transmitted from parents to their offspring. This seminal work was published in 1865 by Gregor Mendel. Crossing garden pea plants and observing traits such as flower color, he noted that the offspring did not get a blended color, but retained one of the traits from either the maternal or paternal parent. Further, if an offspring generation (F1) was generated from two parental lineages, each with consistent but different colors, and F1 subsequently was crossed with itself to create an F2 generation, then he noted that F1 individuals all had the same color but in F2 the colors segregated as 3:1. This led to a postulation that traits are transmitted by heredity factors that appear in pairs, and that each factor is either dominant *A* or recessive *a*. Later these factors were coined *genes* and the alternative variants



they can assume *alleles*. Mendel stated that for each gene an individual receives one allele from each parent resulting in either a homozygous;  $AA$  or  $aa$ ; or heterozygous  $Aa$  genotype. These findings, together with the later identification of chromosomes as the carrier of genetic factors, lay the foundation of classical genetics, and its terminology is still intensely used in modern genetics.

### Mendelian disorders

Mendelian disorders are diseases that follow Mendelian inheritance patterns. Mendelian inheritance imply that a particular allele must have a nearly complete *penetrance*, meaning that it solely can determine the state of the trait. In other words, carriers of a Mendelian risk-allele will almost certainly develop disease. Observing a Mendelian pattern also typically imply that a single defective genetic loci can be identified to transmit the trait, be it in the form of a gene mutation, translocation, copy number variation or any other type of genetic variation. Mendelian diseases are therefore often referred to as monogenic diseases.

These characteristics make genetic mapping by familial linkage analysis a suitable study design to find disease-associated genetic loci. In linkage analysis disease-prone families are collected and polymorphic markers in the genome are used to identify which variants that cosegregate with the disease. Markers closest to the disease gene show the strongest correlation with the disease pattern, and typically the region harbouring a disease gene can be narrowed down to between 100 and several thousand kilobases.<sup>1</sup> Subsequent to the identification of such a candidate region, fine-mapping by positional cloning is applied to identify the specific gene causing the phenotype.

The application of family-based linkage association studies on rare disorders yielded great success during the 80's and 90's and today the genetic basis has been found for about 3000 out of all approximately 6000 rare disorders.<sup>2</sup> The findings are collected in repositories such as the *Online Inheritance In Man* (OMIM) database.<sup>3</sup>

It can be noted that high penetrance leads to a high selective pressure against risk-variants. This reduces the frequency of these variants and has the effect that the disorders are rare. Then how about the common diseases? This leads us to the notion of complex diseases.

### Complex diseases

Mendel's laws were not accepted without controversy. Many traits observed in nature are continuous, or quantitative, rather than binary and discrete, and at a first glance this may appear inconsistent with Mendelian theory. However, in 1918 Ronald A. Fisher resolved this dilemma by showing that multiple Mendelian factors, each with a weak effect, can contribute to the observed variation of an individual trait such that the sum of effects from multiple loci generate a quantitative outcome.<sup>4</sup> Furthermore, the central limit theorem implicates that such a trait will follow a Gaussian distribution in a population, which is similar to what is observed.<sup>5</sup> Today we still have a view that reflects this century-old distinction; human disorders are categorized as either being rare monogenic Mendelian disorders, or

denoted to be non-Mendelian, polygenic, or complex, disorders.<sup>6</sup> The term "complex trait" therefore refers to any phenotype that does not exhibit classic Mendelian inheritance attributable to a single gene locus, and in fact, the most common traits of medical relevance do not follow simple Mendelian genotype to phenotype correspondence, including traits such as heart disease, diabetes and cancer.

## Genetic epidemiology

Genetic epidemiology can provide important clues to the genetic architecture of a disease by examining disease patterns in families and populations. This section introduces three key epidemiological concepts.

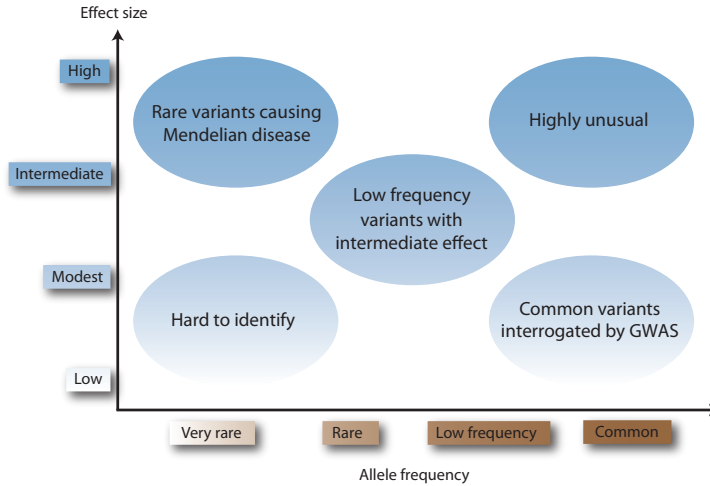
A key concept is the *heritability* of a trait. Formally, the narrow-sense heritability is defined as the proportion of phenotypic variance in a population that can be explained by additive genetic factors<sup>7</sup>). Let  $P, G, E$  denote phenotype, genotype and environment respectively, then,  $P = f(G, E)$  and  $V(P) = V(G) + V(E) + 2Cov(G, E)$ , where  $V$  is the variance and  $Cov$  the covariance. The genetic variance can in turn be divided into three components,  $A, D, I$  including additive and dominance effects of the alleles at a locus, and interactions between loci, such that  $V(G) = V(A) + V(D) + V(I)$ . The definition of narrow-sense heritability can then be written as  $h^2 = \frac{V(A)}{V(P)}$ . Heritability is typically estimated from studying the correlation of a trait between family members. Such familial aggregation of disease can be inflated by shared familial environment, but the estimates tend to agree with those from animal pedigree studies, suggesting that they have a relatively good accuracy.

Another essential measure is *effect size*, which measures the strength of relationship between two variables within a population. For diseases a dichotomous classification is often used where individuals can be assigned as either cases or controls. The association between an allele and the disease can then be described by the odds ratio,  $OR = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)}$ , where  $Y$  denotes case-control status and  $X$  allelic exposure. One should note that the odds ratio estimates relative and not absolute effect size, in the sense that a common allele with low odds ratio may explain a larger part of the phenotypic variability than a rare allele with higher odds ratio.

Finally, the *power*, also known as sensitivity, of a statistical test is the probability that it rejects a false null hypothesis,  $power = 1 - \beta$ , where  $\beta$  is the false negative rate, also referred to as type II error. Plainly speaking, it is the error of failing to observe a difference when in truth there is one. The power is a trade off with the false positive rate (type I error), also referred to as significance level,  $\alpha$ . Power analysis is commonly used to estimate what sample size is needed in a study to make an observation statistically significant given an expected effect size.

## Allelic architecture of genetic disorders

The difference between Mendelian and complex diseases can be illustrated by the allele frequency and effect size (odds ratio) of the risk-variants identified to be associated to a disease, as depicted in Figure 2.1. The large effect sizes of genetic variants



**Figure 2.1** – Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Adapted from Manolio *et al.*<sup>9</sup>.

involved in monogenic conditions, allow them to be detected by classical linkage analysis in family studies. However, variants that increase the risk of disease by less than about fourfold are expected to generate inconsistent linkage evidence.<sup>8</sup> The question arises where the common diseases can be expected to be found in this two-dimensional space and what type of study would be needed to identify their risk variants. Based on increased knowledge of the human genome, and in particular human genetic variation, a hypothesis was made that common diseases are caused by common variants. This hypothesis is described in more detail in the next chapter where genome-wide association studies (GWAS) are introduced. Next, a brief introduction to human genetic variation is given.

## 2.2 Human genetic variation

A prerequisite to achieve deep understanding of the relationship between genotype and phenotype, is knowledge of DNA sequence variation. Extensive variation between individuals has been observed in the form of single nucleotide polymorphisms (SNPs), short insertions and deletions (indels), and larger structural variants such as copy number variations (CNVs) and genomic rearrangements. The public catalogs of genomic variations include approximately thirty million SNPs and six million indels (dbSNP 132),<sup>10</sup> as well as sixteen thousand CNV loci (Database of Genomic Variants v.10).<sup>11</sup> Allele frequencies and the correlation patterns between nearby variants, giving rise to so called haplotype structures, have been mapped by the International HapMap Project across several populations.<sup>12</sup> Recently, the

1000 Genomes Project presented results of the pilot phase, including: 179 low-coverage whole-genome sequenced individuals; two high-coverage sequenced father – mother – child trios; and 697 exome sequenced individuals.<sup>13</sup> They reported the location, allele frequencies and local haplotype structures for variants with allele frequency down to 1%, and showed that low-frequency variants (defined as 0.5% to 5% minor allele frequency), vastly outnumber common variants if considering variants found in a whole population. The contribution of such rare variants to disease is to a large degree unknown. Further, they found that on average, each person carry 50 to 100 variants previously implicated in inherited disorders and 250 to 300 loss-of-function variants in annotated genes. This indicates a redundancy, similar to what has been observed in gene knockout studies of model organisms.<sup>14</sup>

### 2.3 The cancer genome

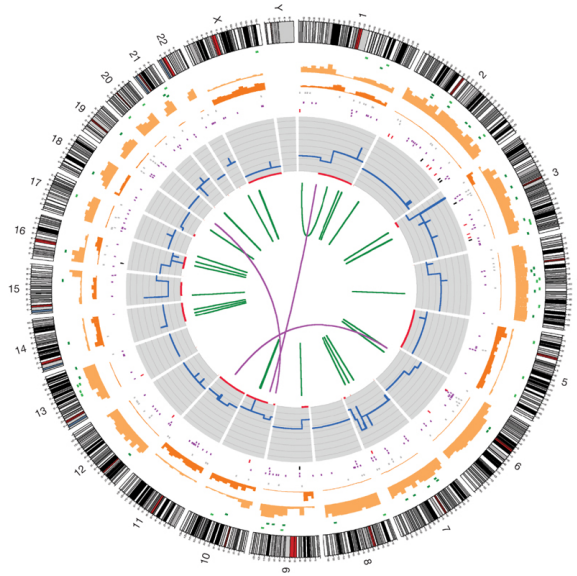
A cancer cell is a direct descendant, through a lineage of mitotic cell divisions, of the fertilized egg and it therefore carries the germline variants inherited from the parents of the patient. However, the genomes of cancers accumulate somatic mutations over the lifetime of an individual, via DNA damaging mutagens or imperfect replication, and can thereby acquire a set of mutations that allows it to proliferate autonomously. Once a cancer cell reach a proliferating state its mutation rate can drastically increase.<sup>15</sup>

A cancer cell therefore contains a huge amount of somatic variation, as illustrated in Figure 2.2. A major task is then to identify which variants that confer a growth advantage to the cell, and are causally implicated in oncogenesis, and which that do not contribute at all. These are termed *driver* and *passenger* mutations, respectively. This is similar to the quest of finding disease-associated loci among germline variants, and large sample sets of cancers are needed to catalog and to identify variants that occur more frequently than by random.<sup>16</sup> A number of common mutations are known, such as in *TP53*, but many genes contributing to cancer development seem to be rare.<sup>17</sup>

### 2.4 Testicular dysgenesis syndrome

The reproductive system exerts an especially important function as it is crucial for a species survival. The recent trend of declining reproductive health in the western world therefore raises special concern. Epidemiological studies of testicular cancer show that during the last 50 years the prevalence has continuously increased, and in particular in Denmark.<sup>19</sup> Similar negative trends for other male reproductive disorders have also been reported, including subfertility, cryptorchidism and hypospadias.<sup>20,21</sup> Notably, these trends are synchronized, following the same ethnic and geographical pattern.

It has also been observed that the aforementioned reproductive disorders seem to be risk factors of each other. In particular cryptorchidism, a congenital malformation where the testis has not fully descended to the scrotum, is a well-established risk factor for testicular cancer.<sup>22</sup> In fact, most of the established risk factors of



**Figure 2.2** – Catalog of the total somatic mutation content of a malignant melanoma cell line. From the inside out, the core displays the structural rearrangements; intrachromosomal are in green, interchromosomal in purple. The next ring out shows the chromosomal copy number in histogram form, with inner red patches indicating regions of loss-of-heterozygosity (LOH). Further out, several rings of single base coding substitutions are shown (black tiles show splice site mutations, red stop-gained, purple non-synonymous and grey synonymous changes). The inner dark orange and outer light orange histograms represent non-coding mutations, relative frequencies of homozygous and heterozygous mutations, respectively. In the final ring before the chromosome indicators, indels are shown in green; light green represents insertions and dark green deletions. Adapted from Pleasence *et al.*<sup>18</sup>.

testicular cancer are related to early life events, including *in utero* exposure of endocrine (hormonal) disruptors,<sup>23</sup> and testicular carcinoma *in situ*, a precursor of testicular cancer that is presumed to be derived from primordial germ cells that escaped normal differentiation during early fetal development.<sup>24</sup> Further, testicular cancer is an early onset cancer, and is the most common malignancy among young men, 15 to 40 years old.

Taken together, these observations led to the proposal that testicular cancer, cryptorchidism, hypospadias and a subset of impaired spermatogenesis all originate from a disturbance in early fetal testis development and that they may share etiology. This hypothesis was formulated in 2001 by Skakkebaek *et al.* and coined the *Testicular Dysgenesis Syndrome* (TDS).<sup>25</sup>

Adding further credibility to the TDS hypothesis is the fact that genetic studies provide emerging evidence that susceptibility factors with pleiotropic effects (effects of the same variant on multiple characteristics or disease risks) are less rare

than previously expected, and genetic links have been shown between not only subtypes of the same disorder but also between apparently disparate conditions.<sup>5</sup> For example, one region of chromosome 8q12 is associated with several forms of cancer, and a growing number of loci are associated with more than one autoimmune disease. This inspires a more a holistic understanding of the origins of human disease, and corroborates the importance of work performed in the area of phenomics, where all human diseases and human disease genes are integrated and jointly analyzed.

### Genetic effects

Testicular germ cell tumors (TGCT) has the third highest heritability among all cancers. The risk of TGCT to brothers and sons of TGCT cases is elevated approximately 6-8 fold and 4-6 fold, respectively,<sup>26</sup> and 37 fold in twins.<sup>27</sup> Large ethnic variations may also indicate a difference in genetic susceptibility. Incidence rates are highest in Northern European populations (8.0-9.0 per 100,000), contrasted by the low rate in Asian and African populations (<1 per 100,000),<sup>28</sup> as illustrated in Figure 2.3.

Despite these evidences of a substantial genetic component to testicular cancer susceptibility, linkage studies have failed to identify any high penetrance loci.<sup>29</sup> Further, candidate gene studies have shown mixed results. The most reliably associated candidate genetic factor is a partial deletion of the azoospermia factor region c (AZFc) of the Y chromosome – known as the *gr/gr* deletion – previously associated to infertility.<sup>30</sup> Recently, three genome-wide association studies on testicular cancer reported five susceptibility loci of high reliability implicating the genes *KITLG*, *SPRY4*, *DRMT1*, *TERT* and *ATF7IP*.<sup>31-33</sup> The effect size for variants in the region of *KITLG* is the highest reported for any malignancy so far, in agreement with the high familial aggregation as compared to other cancers. However, the identified loci account for about 13% of the heritability,<sup>34</sup> indicating that genetic factors remain to be found. Notably, all loci are involved in the biology of primordial germ cells, and specifically implicated in the the *KIT-KITLG* pathway, which regulates the survival, proliferation and migration of primordial germ cells.<sup>35</sup> Further, it is interesting to note that *KITLG* is the most frequently somatically mutated gene in testis tumors,<sup>36</sup> linking the germline genetic variations to the somatic alterations of testicular cancer.

### Environmental effects

Despite the significant heritable component of TGCT, genetic effects have been estimated to account for only about 25% of testicular cancer susceptibility.<sup>37</sup> Further, the recent adverse trend in male reproductive health suggests a critical role for environmental factors.

The effect of drugs and environmental chemicals is a major health concern. In humans, normal development and functioning of the reproductive system of males and females are dependent on carefully regulated hormone levels. The main hormones in reproduction are controlled by three endocrine glands, collectively

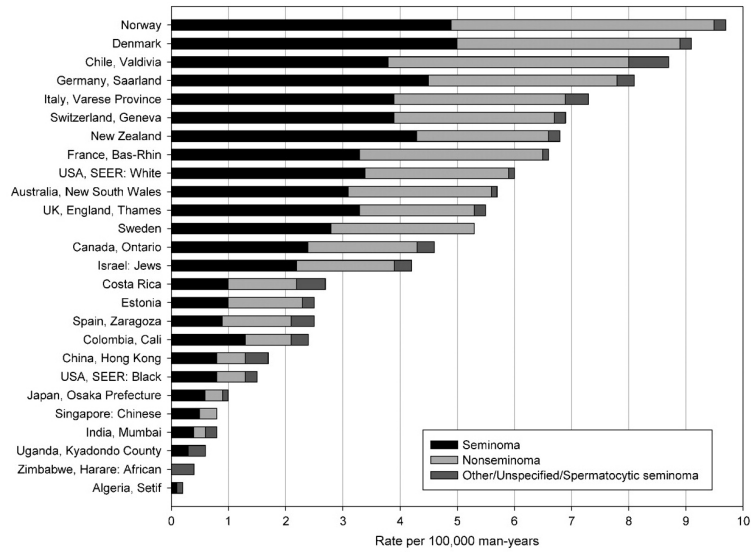


Figure 2.3 – Testicular cancer incidence rates (1998-2002). Adapted from Chia *et al.*<sup>28</sup>.

named the hypothalamic-pituitary-gonadal (HPG) axis. The hypothalamus produces gonadotropin-releasing hormone (GnRH). In response to GnRH the pituitary gland produces luteinizing hormone (LH) and follicle-stimulating hormone (FSH). These two hormones in turn control the production of testosterone, estrogen and inhibin in the gonads. To connect this regulatory system, these hormones form feedback-loops to the hypothalamus.<sup>38</sup>

There is a number of suspected human reproductive toxicants including polychlorinated pesticides, phthalates, bis-phenol A, poly-brominated flame retardants (PBDEs), perfluorinated compounds (PFCs), octyl-/nonyl-phenols and metals.<sup>39</sup> Several of these are endocrine disruptors but other exert their adverse effects through oxidative stress, ionic mimicry and genetic alterations such as mutations or epigenetic modifications. Some studies provide tentative support for an association with phthalates and persistent halogenated compounds to reproductive disorders, but assessment is hindered by lack of a good animal model for testicular germ cell tumors, the time lag between the presumed sensitive period during fetal development and the clinical appearance, and the rarity of the condition.<sup>40</sup>

Finally, it is not unlikely that gene-environment interactions play a crucial role, where certain genetic variants modifies sensitivity to the environment. A recent candidate study of hormone-metabolizing genes suggests that polymorphisms of *CYP1A1* modify the effect of endocrine disruptors, but more evidence is required.<sup>41</sup> For genome-environment-wide association studies to be performed large consortia are likely to be needed. Alternatively, as environmental effects are mediated through intermediate events, such as changes in gene expression, epigenetic mod-

ifications, somatic mutations and interference by small RNAs, inclusion of such external biological knowledge can be of use.





---

## Chapter 3

# Molecular characterization of complex phenotypes

---

Low effect sizes and incomplete heritability of complex diseases present a challenging task when searching for their genetic basis. How can we gain further insight to the molecular basis of conditions with such properties? The work presented in this thesis approaches this from a data-driven perspective, via generation and analysis of system-wide omics data.

This chapter describes the rationale behind, and properties of, the omics data that was used in this thesis, including data from the genome, transcriptome and proteome, as well as highlighting the importance of chemical exposures. Finally, a background to the rationale of systems biology and integration of omics data is presented. More technical aspects regarding experimental platforms and data analysis is presented in chapter 4.

### 3.1 Genome-wide association studies

For most common diseases, linkage analysis has achieved only limited success.<sup>1</sup> Similarly, other types of studies impose other restrictions: admixture mapping requires differences in allele frequency and incidence between two populations, and candidate gene studies are not genome-wide nor agnostic. With the completion of the first draft of the human genome in 2001 the first step towards genome-wide association studies that scan the whole genome for genetic risk factors was taken. In association studies a genetic variant is genotyped in a population for which phenotypic information is available. If a correlation is observed between genotype and phenotype, there is said to be an association between the variant and the trait.

Which variants should then be targeted for genotyping? As mentioned above, the databases currently contains about 30 million human SNPs. Before sequencing

becomes cheaper this is a prohibitively large number of variants to be genotyped on a large set of samples.

### Common disease – common variant

The allele frequency of variants that underlie common diseases has been under debate for at least a decade.<sup>2</sup> Estimations of the frequency spectrum of all single nucleotide polymorphisms made around year 2000 indicated that most of the genetic variation between any two individuals are accounted for by common variants with frequencies >5%.<sup>3</sup> Assuming that variants underlying common disease have a similar frequency spectrum to that of all variants, would imply that the major part of the heritability can be accounted for by common variants. This is what the common disease – common variant (CDCV) hypothesis proposes.<sup>4</sup> The hypothesis is plausible if the estimated frequency spectrum is accurate, and there has been no intense negative selection of disease susceptibility alleles. There are several arguments to support a low negative selection. 1) A modest effect size, where the joint effect of many variants result in a phenotype, would be in line with the quantitative character of complex traits. 2) The late onset of many common diseases can make the variants elude selection as high ages historically were more rare. 3) Alleles that provide resistance to infectious diseases are under positive or balancing selection and can therefore reach relatively high population frequencies. An example is the heterozygotic genotype that is protective against malaria, whereas a homozygote develops sickle-cell anemia. 4) Alleles selected due to a historical advantage, for example for fat storage, may not be beneficial in a modern environment, also known as the thrifty gene hypothesis.<sup>5</sup>

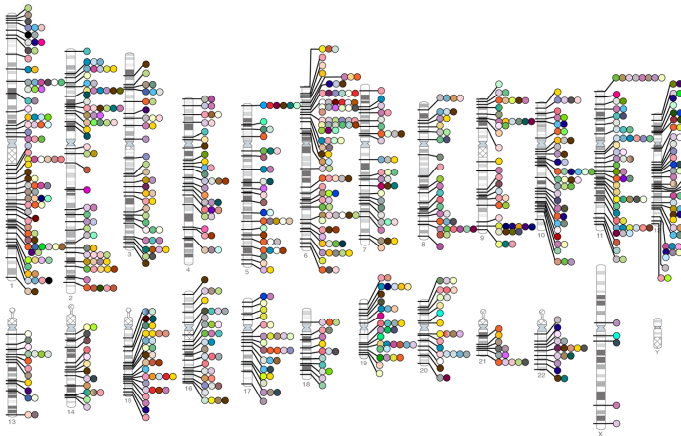
The alternative to CDCV is the classical disease heterogeneity hypothesis (or multiple rare-variant hypothesis), where there is a wide range of disease alleles, each relatively rare.

### HapMap and linkage disequilibrium

To perform an association study of common variants, we need to know all the common variants in the population from which the cases and controls are sampled. To this end, the International HapMap Project was initiated, providing a catalogue of common variants in a range of populations.<sup>6</sup> It demonstrated that site variants cluster in local neighborhoods, haplotypes, of strongly correlated variants. Therefore, representative SNPs can be chosen to tag a set of variants within a haplotype, as it can predict the alleles of correlated neighbouring SNPs. This phenomenon of non-random association of alleles at different loci is termed linkage disequilibrium (LD). Using the LD-patterns determined by the HapMap project it was shown that the selection of a few hundred thousand SNPs are sufficient to cover most of the common variation in the genome.

### The GWAS explosion

At the time of completion of the first phase of the HapMap project in 2005, the rapid progression of SNP genotyping microarrays had made it possible to simultaneously



**Figure 3.1** – Published genome-wide association studies<sup>8</sup>. Each circle indicates an associated locus and colors different types of diseases.

interrogate hundreds of thousands of SNPs on a single microarray. This put the last piece in place to enable the execution of genome-wide association studies of common variants. In 2005 the first proof-of-principle study was published on age-related macular degeneration identifying a SNP with a relatively high effect size that for the first time related inflammation to macular degeneration.<sup>7</sup> Since then we have witnessed an explosion of genome-wide association studies, and today more than 900 studies can be found in the NHGRI catalog of published GWAS, as depicted in Figure 3.1.

Two trends are evident in the published GWAS data. First, with few exceptions the regions implicated by GWAS were not previously known. Candidate genes based on prior knowledge of pathophysiology or intuition have usually not been identified. Second, the majority of these findings were not in the coding region of a gene. This complicates the understanding of how a locus contributes to disease, and further sources of information may be needed to guide their interpretation.

## 3.2 Transcriptomics

Genetic variability is an underlying driver of inter-individual phenotypic variability. However, at an intermediate level, a genetic variation can induce differences at the transcriptional and translational level. Identification of candidate genes can therefore be accomplished by associating transcriptional expression levels with phenotypic traits. The power to identify molecular entities that covary with a given trait is greater than compared to genome-wide association studies, due to the lower number of entities tested and the dynamic range of transcript abundance. On the other hand, expression levels are modulated by external environmental actions,<sup>9</sup> they are highly susceptible to cell-population heterogeneity, and genes are involved

in multiple interactions within a cell, making it difficult to discern causal changes from downstream effects. This is discussed in somewhat further detail below.

### A snapshot of the cellular state

Transcriptional profiling offers the possibility to investigate dynamic aspects, for instance, response to external factors and time-dependent processes such as the cell-cycle. It also permits the identification of tissue-specific features. This stands in contrast to genomic analysis, as the DNA within an organism is relatively stable over time. In paper III we exploited the tissue-specific and dynamic property of transcription data by complementing genetic data with mRNA samples taken at different time-points from developing testis.

The sensitivity of transcriptomes to environmental conditions not only offers opportunities but can also be problematic as it can be difficult to control. Many expression studies of cancer use material from tumor biopsies, which normally contain a multitude of different cell-types, each with their own transcriptional signature. To isolate the cell-population of interest we therefore performed laser-microdissection of labeled testicular *carcinoma in situ* cells as described in the first paper presented in this thesis.

Another hurdle is the difficulty to discern driving changes from secondary "ripple" effects that appear due to signaling pathways and regulatory control within a cell. However, if a differentially expressed gene also have been implicated in complementary studies such as linkage analysis or genome-wide association studies it is more likely to be causative. Such reasoning forms the rationale for integrating genetic and expression data in disease genetic studies.

## 3.3 Protein interactomics

The human genome provides an informative 'parts list' of protein-coding genes, but understanding of cellular systems may benefit from assembling the parts, and in fact, proteins rarely function alone, but rather tend to act in concert with other proteins to accomplish vital cellular functions. This could be in rigid extra- or intra-cellular supportive structures, in relatively stable assembled molecular machines, such as the ribosome, or in transiently formed constellations, such as during the cell cycle<sup>10</sup> or in phosphorylation pathways.<sup>11</sup> These modules, or complexes, are made up of proteins that physically bind to each other. The universal extent of such protein-protein interactions has been well-documented by high-throughput screens in model organisms, such as *C. cerevisiae*,<sup>12</sup> *C. elegans*<sup>13</sup> and *D. melanogaster*,<sup>14</sup> as well as in *H. sapiens*.<sup>15</sup> This has given rise to the term interactome, referring to the known set of protein interactions within a species. Figure 3.2 illustrates the human interactome and three subnetworks within it.

It has become clear that the interactions between proteins are vital for cellular processes to be functional. For example, the impact on human disorders of edge-specific perturbations, where only specific interaction surfaces of a protein are disrupted, were recently shown.<sup>16</sup> Due to this modularity, where small networks of proteins perform a function, protein-protein interaction data has proved

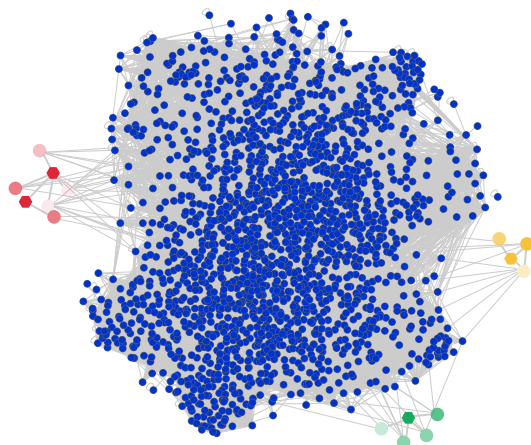


Figure 3.2 – The human interactome. Nodes represents proteins and edges interactions between proteins.

to be an important resource for the prediction of disease genes and phenotypic effects of gene mutations.<sup>17–20</sup> These interaction networks also carry the potential to explain comorbidity,<sup>21</sup> the partial phenotypic overlaps observed between different diseases,<sup>22</sup> and pleiotropic effects.<sup>23</sup>

There are however reasons to practice caution, since protein-protein interaction information can be error-prone,<sup>24,25</sup> and an observation that two proteins can interact, does not necessarily imply that they do interact *in vivo*. Further, in contrast to the genome sequence, the connectivity map is dynamic and spatiotemporally dependent, much in the same way as the transcriptome. For example, many interactions that control cellular behavior are dependent on post-translational modifications. Basically for every cell type, stressor, stimulus, nutritional condition, or other parameter of interest, there is a unique interactome. In this thesis I make use of the interactome compiled and maintained by our lab, and it is considered static. A short primer on the two main experimental techniques that are used to generate this type of data, as well as the issues of quality control, and compilation of a complete interactome from the available data resources, are further discussed in chapter 4.

### 3.4 Chemical biology

Genetics rely on the assumption that we can explain an observed phenotype based on an organism's genetic architecture, however, this raises a fundamental question. To what extent can the cause of a disease be explained by genetics alone?

Organisms constantly respond to and interact with the environment via small molecules, such as via inhalation, orally, or dermally. Internally, they are essential parts of the metabolic and endocrinological pathways. Environmental factors,

or small molecules in the form of drugs, can therefore cause perturbations of a biological system, or reverse a pathological state. Concordantly, many environmental risk factors for complex human diseases have been identified by epidemiological studies. It is however a demanding task to unravel how genotypes at specific loci modulate individual responses to environmental risk factors, so called gene-environment interactions, and these effects are largely unknown.<sup>26</sup> This has spurred a recent development where traditionally relatively separate fields such as genetics and toxicology have been merged together as to study the external modification of biological systems by chemical compounds in a more integrative manner, giving rise to scientific disciplines such as chemical biology, pharmacogenomics and toxicogenomics. One of the goals is to provide patient-group stratified, or even personalized, medical treatments based on the genetic profile and individual response to a specific drug.<sup>27,28</sup>

### Chemical network biology

A recent review questions the traditional drug discovery approach – one drug, one target, one disease – and if the assumption that one drug only interacts with one target holds true. High attrition rates may be a result of that drugs designed to be highly selective are not the most efficient or cause less side-effects.<sup>29</sup> Notably, one study that investigated drug-targets in a human protein-protein interaction context found a recent trend toward the development of new compounds directly targeted at disease gene products, whereas previous drugs, often found by trial and error, appear to target proteins only indirectly related to the actual disease molecular mechanisms.<sup>30</sup> These findings raise the question whether targeting network properties may have a larger effect on a biological system rather than directly targeting the product of a single mutated gene.

Corroborating the idea that several components of a network need to be affected to cause a discernible effect is the observation that biological systems seem to have a built-in redundancy, making them more robust and fault-tolerant.<sup>31</sup> Biological networks are known to have scale-free properties: a small number of highly connected hub-proteins, and a large number of proteins with only a few interactions. Targeting a hub-protein can therefore have a severe effect on the system, but for the majority of proteins the function can be recovered by other parts of the network.

This may explain the remarkable inertia that model organisms have been shown to possess against gene knock-outs. The proportion of essential genes has been estimated to reach 19% in several organisms, for example, in yeast as few as 15% of all systematically knocked-out genes led to fitness defects.<sup>32</sup> Possibly a multitude of proteins may therefore need to be perturbed before a functional module is disrupted. In line with this growing body of knowledge on the importance of regulatory networks, this thesis work describes several projects where network or pathway based analysis is used.

During the course of my PhD I have performed work relating to the effect of small molecules in protein networks.<sup>33</sup> The study was excluded from this thesis to limit its scope, but it is however clear that small molecules play a critical role in disease development and treatment.

### 3.5 Integrative systems biology

Systems biology presupposes that no life form can be imagined without the complex systems formed by interacting macromolecules and metabolites, cells and organs.<sup>34</sup> Considered as separate entities, the sheer number of genes, proteins or cells, seem to fail to account for the differences in complexity between organisms. Despite a gene list not much longer than that of a round worm, evolution seem to have shaped humans into something considerably more sophisticated. This intuitively suggests that interconnections between macromolecules, both at local and global levels, might be able to generate system properties or behaviors that are fundamental to life.

#### More than the sum of its parts

New properties emerge as the components of a system are integrated in networks and coordinated in time and space. The first evidences of this already appeared half a century ago, when Delbrück provided a theoretical model of how a positive feedback loop with two molecular entities can form a bistable switch, where infinitesimal perturbations at a specific time point can lead to either an activated stable "on state" or an "off state" that remains stably off (chaotic behavior). Not much later such chaotic bistability was empirically demonstrated for the *lacZ* operon.<sup>35</sup> Similarly, negative feedback loops have been shown to be able to generate homeostatic behavior, where a system is robust against variable input, and to generate oscillatory phenomena.<sup>36</sup>

Modelling of dynamic systems remains an active field in contemporary science, and recent examples underlines the importance of regulatory circuits in cell cycle regulation, signaling and circadian oscillators. Relatedly, the new field of synthetic biology where biological systems and circuits can be *de novo* engineered offers exciting opportunities to study dynamic properties.

#### Systematic mapping and tsunamis of data

Systematic high-throughput approaches often provide essential data that enable the investigation of system properties. Techniques such as sequencing and proteomics have rendered informative lists of genes and proteins, and the mapping of interconnections between molecules are well underway. Gene-gene, gene-protein and protein-protein networks can now be studied thanks to the application of experimental methods such as combinatorial knockouts, chromatin immunoprecipitation, and yeast two-hybrid. Quantitative measurements of complete transcriptomes produce information that can help elucidating transcriptional networks. Next generation sequencing in conjunction with databases containing pathways, reactions, and functions of classes of enzymes, can help in reconstructing and understanding the metabolic networks of unknown bacteria and parasites.<sup>37</sup> Imaging methodologies can provide spatiotemporal data, as an example, such techniques was recently used to make a video of the 24 first hours of development of an entire zebrafish embryo and was thereby for the first time able to exactly reveal what



goes wrong in the embryogenesis of a mutant strain.<sup>38</sup> Further, patient records remain a to a large extent unused resource, and the corpus of scientific literature continuously grows.

### **Data driven hypothesis and knowledge generation**

The 'tsunami' of data that novel high-throughput techniques present us with introduces unprecedented challenges. To abstract representable knowledge from the masses of data, and to understand the properties of and information flows of networks, new tools are continuously emerging.

Having mapped the major set of parts of a specific biological feature, such as single nucleotide variations, the complete list of parts can be used in screening designs. The basic idea behind such system-wide measurements is to be hypothesis-free, and let the data direct the scientist to an unbiased finding. Being system-wide, a feature shared by these screen-based studies is that the number of measurement points (features) are huge, often in the range of thousands and millions. This is an inherent complication of these methods, and especially as the number of samples for which these measurements were made are often orders of magnitudes lower. It presents a problem as the aim of these experiments are often to identify which among the thousand of features that carry information regarding a certain trait of the samples.

### **Data integration to boost statistical power**

One obvious means to facilitate identification of informative features in a vast search space is to increase the amount of information. As there often exists prior knowledge about disease pathobiology, or previous experimental data, we attempt to utilize these sources of information to detect candidate biological entities that affect disease development and progression. In several of the papers described in this thesis, the primary data received from our collaborators has been augmented with previously published large-scale datasets. The usefulness of integrating evidence from multiple data types, such as genomic variation, gene expression and proteomics to extend our understanding of diseases have been reported in a number of recent publications.<sup>39-42</sup>

---

## Chapter 4

# Technical and analytical aspects

---

This chapter describes experimental platforms and provides a more technical background to the data analysis performed in the studies presented in part II of the thesis.

### 4.1 Microarrays

The first microarrays that allowed simultaneous assessment of thousand of transcripts appeared in the mid-1990s, and their development was driven by the availability of almost complete catalogs of expressed transcripts in human and model organisms. Transcriptional profiling by microarrays has since, among other things, been used to elucidate gene function, in clinical diagnosis and in the exploration of possible causes and mechanisms of disease. After their establishment as useful tools for gene expression profiling, microarrays have found new areas of application, including comparative genomic hybridization (CGH) and arrays that can interrogate SNPs and CNVs.

Microarray studies have a number of steps in common, these include sample preparation, hybridization and preprocessing. Once a preprocessed and quality controlled data set is at hand, there is a range of types of analysis that can be performed, including: univariate ranking of genes or SNPs, network or module inference, functional annotation and enrichment analysis, meta analysis, and integrative analysis.

### Gene expression profiling

Gene expression microarrays contain thousands of oligonucleotides, or probes, that hybridize to their complementary target RNA (or DNA) which has been generated from a biological sample. Typically the target RNA have been generated through

amplification by qPCR, selected for protein coding genes by poly-adenylate tail capturing, and labeled by fluorophores. Following hybridization, unbound targets are washed away and the target-probe intensities are detected by fluorescence scanning.

### Preprocessing

Raw intensity measures obtained are not directly comparable, neither between probes nor between arrays, due to technical and experimental variation. Preprocessing of expression arrays typically involves three steps: (1) background correction to account for background noise such as spatial intensity trends on a chip, and differences in probe affinities; (2) normalization to correct for between-array variation; and (3), probe set summarization, in order to retrieve one value for several probes targeting the same region. These preprocessing steps are today to a large extent standardized for conventional 3' expression arrays and commonly used software packages are available in R Bioconductor.<sup>1</sup>

### Univariate ranking

After preprocessing, and quality control based on summary statistics and visualizations to remove sample and probe outliers, most experimental designs will demand the generation of ranked lists of genes. Typically the rank is based on the differential expression observed for a transcript as a consequence of varying sample treatments or sample conditions used in the experiment. Such analysis often tests for difference in mean expression between subsets of samples, and, using assumptions of normal distributions, parametric tests such as t- or F-test like procedures can be used. As it has been observed that using conventional t- or F-tests result in a large portion of false positives due to poor estimates of the expression variance, recent methods such as eBayes and SAM have been developed that improve the estimates of the variance by instead using a standard error estimate that is obtained by borrowing information from the whole ensemble of genes.<sup>2,3</sup> Differentially expressed genes that are to be followed up in downstream computational or experimental analysis can then be chosen based on the resulting p-value. For downstream computational analysis a less conservative cutoff than the conventional 5% significance level may be suitable, or the complete ranked list could be utilized such as to test for annotation skewness towards high-ranked genes. A more detailed description of such downstream analysis is presented in the analysis section below.

### Comparative genomic hybridization

In contrast to expression profiling that is concerned with the abundance of transcripts, comparative genomic hybridization (CGH) is concerned with the abundance of actual DNA. Specifically, the goal of CGH is to determine the copy number of genetic regions. Array CGH (*aCGH*) is simply a CGH measurement performed with a microarray and the experimental procedure is similar to that of expression profiling except that the starting material is DNA instead of RNA. A difference is that as the aberrated genetic regions may be large. There also exist arrays based

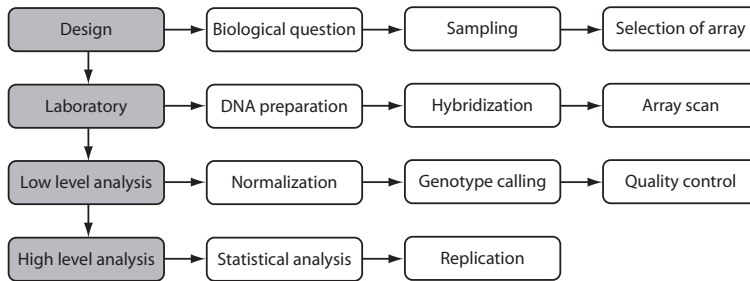


Figure 4.1 – Typical workflow of a genome-wide association study.

on bacterial artificial chromosomes (BAC) instead of oligonucleotides. aCGH has primarily been applied in cancer biology as an important aspect of cancer cells is their tendency to gain or lose large proportions of their genome. Recently, the use of aCGH has become somewhat replaced by the SNP genotyping platforms.

### Genome-wide association

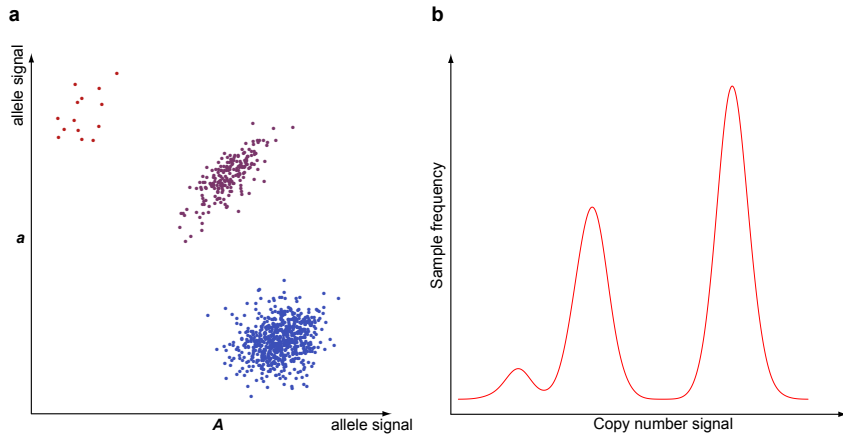
The typical workflow of a genome-wide association study (GWAS) is illustrated in Figure 4.1. It includes sampling of a study cohort, laboratory work, low level analysis that involves genotyping and quality control, and high level analysis to identify loci associated to the trait of interest. Below a brief account is given of the DNA microarray used, quality control and statistical testing.

#### The genotyping platform

The GWAS presented in this thesis used the Affymetrix Genome-Wide Human SNP Array 6.0, which is a 'hybrid' genotyping platform containing 906,600 probesets targeting SNPs as well as 946,000 'copy-number' probes – non-polymorphic probes that are optimized for copy-number measurement and unconstrained by the locations of SNPs. Each SNP is targeted by a probeset of three or four 25-mer probes for each of the two alleles, whereas there is only one replicate of the copy-number probes.<sup>4</sup> An important distinction is that by design the SNP probesets target common genetic variation, whereas analysis of CNVs allows identification of novel rare CNVs.

#### Quality control

After fluorescence scanning and normalization of the intensity signals, a genotype of each SNP in each sample is assigned. The genotype is typically inferred with methods involving an expectation-maximization algorithm, which determines which of three possible genotype clusters a SNP belongs to. Figure 4.2 illustrates how a scatter plot of the intensities for a given SNP forms three separate clusters, corresponding to the genotypes of the samples.



**Figure 4.2** – Example of intensity signal of all samples at a locus. (a) Scatter plot of allelic signal for a SNP. (b) Smoothed histogram of copy number signal.

Once genotype calls have been obtained, a set of quality control steps are applied to determine the quality of genotypes and samples, and those of low quality are excluded. The quality control steps of our GWAS is illustrated in Figure 4.3. Genotype quality was assessed by compliance to Hardy-Weinberg equilibrium and confidence of genotype (cluster) assignment. SNPs with low minor allele frequency were excluded. Further, when an association for a SNP has been identified, it is important to do a visual inspection of its signal intensity plot, to ensure the genotyping quality. Sample quality control involved evaluation with regard to ethnicity, familial relationships, inbreeding and high fraction of failed genotype calls.

### Univariate association testing

As opposed to the continuous nature of expression levels, genotyping render discrete distributions with a domain of only three possible values. In a case-control setting a  $2 \times 3$  or  $2 \times 2$  contingency table can be used to depict and analyse the case and control frequency distributions of a SNP. Several tests can be used to assess the association of a SNP to the trait. An allelic test compare the allelic distributions by converting the genotype frequencies into allelic counts (A, a). One can also perform a genotype-based test. A typical test statistic used in single-marker analysis for case-control studies is the Cochran-Armitage trend test (CATT), derived under the assumption of an additive mode of inheritance. Analysis of the power of this test has shown that it has a relatively high power to detect associations regardless of the underlying genetic model.<sup>5</sup> An alternative is the MAX test, which evaluates three possible genetic models (additive, dominant and recessive), and it has been shown to be advantageous to CATT as it does not assume an additive model,<sup>6</sup> and it was used in the GWAS presented in this thesis.

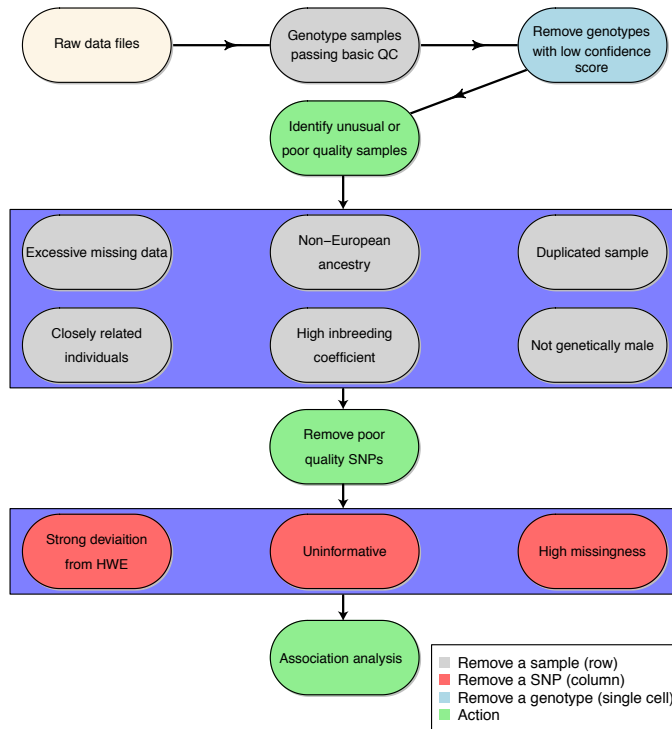


Figure 4.3 – Quality control steps performed in the GWAS presented in paper III.

### Multiple testing correction

Multiple testing correction to find SNPs of genome-wide significance is of special importance in GWAS due to the vast number of SNPs. Given a high number of features in relation to samples it is not unlikely that a few of the features will correlate with a certain sample characteristic even if they in reality are independent. This can be illustrated by the fact that if you let one million people throw a six-sided dice four times it is not unlikely that you find several persons having four sixes. Clearly, a single observation of four sixes among such a large number of persons can not make you infer with any reliability that that person's dice is unbalanced. In statistical terms this phenomena is referred to as *multiple testing*.

Let  $V$  be the total number of type I errors from a family of tests. The family-wise error rate ( $FWER$ ) is the probability that we observe at least on type I error,  $\alpha_{FWER} = P(V \geq 1) = 1 - P(V = 0)$ . The probability that no errors is observed given  $n$  independent tests is  $P(V = 0) = (1 - \alpha_{singletest})^n$ , and therefore,  $\alpha_{FWER} = 1 - (1 - \alpha_{singletest})^n$ . Further, it can be shown that  $\alpha_{FWER} \leq n * \alpha_{singletest}$  even if we don't assume independent tests. This multiplication by the number of tests as to control the FWER is called *Bonferroni correction*.

Bonferroni correction is very conservative, in the sense that it provides a strong control of the false positive rate, but at the cost of power. For large-scale multiple testing, control of the false discovery rate *FDR* is often preferred, defined as the expected proportion of false positives  $\alpha_{FDR} = E[\frac{V}{R}]$ ,  $R = V + S$ , where  $S$  is the number of true positives.

In genomic applications, such as GWAS and transcriptomics, the tests are seldom independent as genes and SNPs are correlated. To retain the distribution characteristics empirical methods are commonly used, typically by permuting the case-control status and resampling for a large amount of permutations to generate a null-distribution.

### Marker validation

Due to the high number of tests, GWASs are relatively sensitive to technical biases (such as batch effects) and biological variation (such as population stratification), which can generate false associations.<sup>7</sup> GWAS designs therefore typically use a two-stage approach where candidates from a discovery stage are put forward to a replication stage with an independent validation cohort. Since the replication phase is typically performed for a small set of selected markers, a different genotyping assay is used, and thereby reducing the risk of findings due to technical artifacts.

### CNV association

CNV and SNP array analysis differ due to three main reasons: (1) CNVs often have a varying number of alleles, where copy-numbers typically vary in the range 0-4 copies, (2) a CNV span a genomic region of multiple contiguous loci, and (3) rare and novel CNVs can be detected. CNV analysis typically involve a three-tier approach.<sup>8</sup> Briefly, first-tier methods provide probe-specific estimates of copy number. A common approach employed is to estimate the ratio or log ratio of the intensities at each loci relative to a reference, but absolute quantitation of the allele-specific copy number dosage can also be used. Second-tier algorithms smooth the probe-specific estimates as a function of the genomic physical position to identify alterations spanning multiple loci. This includes segmentation algorithms, regression-based smoothing methods, and hidden Markov models (HMM). Recent methods include both inferred SNP genotype calls as well as first tier intensity signals in the determination of alteration status.<sup>9-11</sup> Finally, third-tier methods perform association of detected alterations with the phenotypic traits of a study cohort. The association may either simply be performed by tests on contingency tables containing the frequency of CNVs,<sup>12</sup> or more sophisticated methods can be used that propagate the uncertainty in copy number state inference.<sup>13</sup> Rare CNVs can be evaluated via association analysis by aggregating their effect on genes, networks or functional modules.<sup>14</sup>

For the analysis of common CNVs the second tier is not a requirement and one can instead use a genomic map of previously identified common CNVs. It is worth noting that the association of CNVs are more difficult than for SNPs, partly due to that discovery and genotyping of CNVs have not been based on separate

technologies. There is however an increasing amount of sequencing projects which will provide improved catalogs of structural variations.

Quality control typically involve the majority of methods used for SNP GWAS, but also include evaluation of CNV confidence and CNV length, as well as sample control with respect to the number and total length of the CNVs detected in an individual.

## 4.2 Protein interactomics

### Experimental detection

The two major high-throughput techniques to detect protein-protein interactions are yeast two-hybrid (*Y2H*) and tandem affinity purification in conjunction with mass-spectrometry (*TAP-MS*). *Y2H* utilizes the fact that a transcription factor has two domains, a DNA-binding domain (BD), and an activation domain (AD) that initiate expression. In order to detect protein-protein interactions the transcription factor is split such that the BD is fused to a protein ('bait') and the AD to another protein ('prey'). If the bait and prey interacts, the transcription factor can bind to its DNA binding site and activate transcription. Typically a transcription factor of common reporter genes are used to detect the presence of transcription, such as the *GAL4-lacZ* system. In *TAP-MS* the bait protein is fused with two 'tag' peptides, protein A, and calmodulin, which are used to purify the bait and its associated prey proteins through affinity purification. The proteins forming a complex are subsequently identified by mass spectrometry. As the precise physical interactions among the proteins in a complex is unknown, data is reported using either a *spoke* model, with one central protein and its interactions, or a *matrix* model where one assume interactions between all proteins of a complex. We prefer the spoke model when possible as it is more conservative and results in fewer false positives.<sup>15</sup> This is not an issue in *Y2H* experiments as it detects binary interactions.

### Inweb: a human interactome

Protein-protein interaction data is available in large database repositories, where data has been deposited or collected via text mining or manual curation of the literature. At our lab, a human interactome has been created, named *Inweb*, by integrating data from a large number of such protein interaction databases, including MINT, BIND, GRID, HPRD, IntAct, DIP, PDZbase, Reactome, and KEGG.<sup>16-24</sup> One unique feature of *Inweb* is that orthologous interactions, *interologs*, has been inferred from model organisms. The level of plasticity and rewiring of protein interaction networks during the course of evolution is relatively low, for example, it was recently estimated that 0.17% of all edges has been rewired between yeast and human.<sup>25</sup>

Quality control is critical as the experimental procedures are known to cause a relatively high number of false positives.<sup>26</sup> To this end a confidence scoring procedure was devised where each protein-protein interaction is assigned a score based on the local network topology and the amount of experimental evidence. This confidence score has been shown to correlate with true interactions,<sup>27</sup> and has been



internally evaluated at our lab by comparisons against benchmark sets of high-confidence interactions.

### 4.3 Digital gene expression

Gene expression profiling by microarray technologies is based on labeling RNA with a fluorophore and then determining its concentration based on the strength of the fluorescent signal. An alternative is to count each observation of a transcript. This approach, also denoted as *digital gene expression*, can be realised by the application of sequencing as the relative abundance of a transcript can be estimated from the number of observed sequences from that gene. Expressed sequence tag (*EST*) sequencing first appeared in 1991 when Adams *et al.*<sup>28</sup> performed Sanger sequencing of clones from brain cDNA libraries. A few years later Serial Analysis of Gene Expression (*SAGE*) was developed that permitted a 10- to 20-fold increase in efficiency by concatenating several short restriction enzyme derived tags into one DNA sequence that is subsequently cloned and sequenced.<sup>29</sup> The next advance in digital gene expression was the development of massively parallel sequencing. The first method described was massively parallel signature sequencing (*MPSS*),<sup>30</sup> where millions of microbeads, each coated with a PCR-amplified cDNA, are packed into a flow chamber. A fluorescence based signature sequencing is then performed by using a combination of restriction enzymes, adaptors and hybridization probes. MPSS has now been superseded by methods that use sequence-by-synthesis, but they are based on the same general approach. The use of sequencing for expression profiling has been termed RNA-Seq.

An important aspect of RNA-Seq is that it does not require prior knowledge of the transcriptome, whereas microarrays are designed based on available genomic and transcriptomic data. This feature enables sequencing to detect novel transcripts, and it can be applied to organisms with an unsequenced genome. Transcriptomic sequencing can therefore provide a low-cost alternative to whole genome sequencing, also called a "a poor man's genome", which in particular can be an option for organisms with large genome sizes, such as for plants that often are polyploid and have undergone extensive retrotransposon amplification.

Several studies have compared RNA-Seq and gene expression arrays and have reported a high degree of correspondence and a higher detection rate in RNA-seq.<sup>31,32</sup>

Although powerful, RNA-Seq is not without challenges, including read quality assessment, read-mapping and *de novo* assembly errors,<sup>33</sup> transcript length biases,<sup>34</sup> and less mature statistical algorithms for the data analysis. A critical issue is the need of sufficient coverage, to cover each locus at a sufficient depth. For transcriptomics this is of particular importance to reach high sensitivity, as a relatively small number of highly expressed loci can account for the majority of the reads in the study. Despite a continuously dropping cost of sequencing, arrays may not become completely obsolete, and could for example be used for capturing specific subsets of the transcriptome that one wishes to target with high sensitivity.

## 4.4 Analysis

A number of general data analysis methods appear in several of the studies presented in this thesis, and the most central ones are put into context below.

### Unsupervised visualization

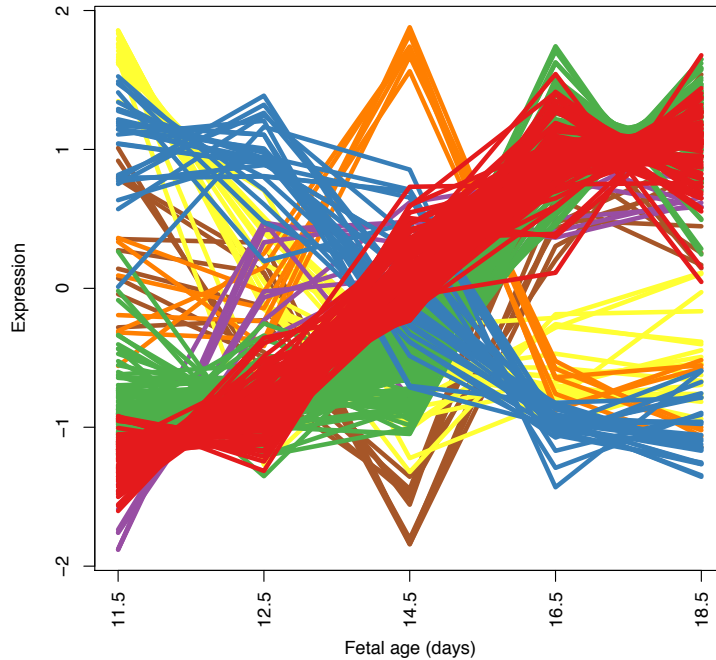
Initial exploration of data generated by arrays commonly involve the use of unsupervised methods to summarize and visualize the complete set of data. A frequently used tool for visualization are dimensionality reduction techniques that map the data into a lower-dimensional space retaining as much information as possible. Common algorithms include principal component analysis (PCA), multi-dimensional scaling and self-organizing maps. A second effective type of methods for visualization are clustering, such as hierarchical clustering methods.

### Clustering and network inference

To identify sets of objects that are similar, commonly defined as being close to each other in a  $n$ -dimensional space based on a distance metric of choice, clustering techniques can be applied.

A widely used clustering method is the *k-means* algorithm which aims to partition  $n$  observations into  $k$  clusters, such that each observation belong to the cluster with the nearest mean. After initial random assignment of cluster means two steps are iterated until convergence: 1) observations are assigned as members of the cluster with the nearest mean, and 2) the mean of each cluster is recalculated based on its members. A similar method, which can be more robust than  $k$ -means, is the  $k$ -medoids algorithm, where the medoid instead of the centroid is used as the prototype for the cluster. The medoid is a point that minimize the sum of distances to a set of data points, just as the geometric median, but the difference is that the medoid needs to be a data point from the data set. The  $k$ -means and  $k$ -medoids can be seen as special cases of the expectation maximization algorithm (EM algorithm), where cluster assignment (E-step) and estimation of model parameters based on the given assignments (M-step) are iteratively performed until convergence. The EM algorithm maintains probabilistic assignments to clusters, instead of deterministic assignments, and use multivariate Gaussian distributions instead of means. The EM is thus able to accommodate clusters of variable size much better, as well as non-spherical clusters that arise from correlation structures. In this thesis the partitioning around medoids (PAM)<sup>35</sup> implementation of the  $k$ -medoids algorithm was applied in paper II and V to identify clusters of genes with similar expression profiles, and the EM algorithm was used in paper III and IV for allelic determination of SNPs and CNVs respectively. Figure 4.4 displays an example of expression profiles from the developing fetal testis clustered by PAM.

One disadvantage of these algorithms is that the number of clusters,  $k$ , needs to be specified by the user. There are several approaches to help determine  $k$ , such as calculating the likelihood for a given  $k$  and use the Akaike or Bayesian information criteria (AIC, BIC) which penalizes for the number of clusters, make use of the Dunn



**Figure 4.4** – Gene expression profile across five time points from the fetal testis of mouse. Colors indicate cluster membership based on a PAM clustering ( $k = 7$ ) of the 250 most differentially expressed genes with respect to time.

index or average silhouette which provide combined measures of compactness of clusters and their separability, or perform cross-validation or perturbation of the data. In paper IV a method is utilized that employs the BIC to determine the number of CNV alleles.

A more generalized type of clustering is subspace clustering, which can identify clusters that only correlate within a subset of samples. This is implemented by for instance biclustering algorithms which simultaneously cluster both rows and columns of a data matrix.

Based on studies of genetic interactions, co-expression, or protein interactions, system-wide genetic, transcription or protein networks can be formed. There also exist resources that have combined several types of evidence of interactions to create one functional network, such as the STRING network.<sup>36</sup> To identify sub-networks of interest within such networks graph-theoretical approaches are typically applied which can assess local topological properties.

## Pathway analysis

As described in the previous chapter macromolecules act in the context of networks. Apart from integrative approaches where knowledge-based frameworks incorporate a multitude of data types somewhat more traditional pathway based analysis has been applied throughout this thesis. Such analysis typically involve annotation enrichment among genes ranked for their association to a condition.

A range of databases contain predefined subnetworks (pathways). During the course of my thesis I collected such gene and protein annotation data from KEGG (Kyoto Encyclopedia of Genes and Genomes),<sup>24</sup> Reactome,<sup>23</sup> BioCarta (<http://www.biocarta.com>), PID (Pathway Interaction Database),<sup>37</sup> GO (Gene Ontology),<sup>38</sup> COSMIC (Catalogue of Somatic Mutations In Cancer),<sup>39</sup> Cyclebase,<sup>40</sup> Inweb,<sup>41</sup> CTD (The Comparative Toxicogenomics Database, <http://ctd.mdibl.org>), OMIM,<sup>42</sup> MGI (Mouse Genomi Informatics database, <http://www.informatics.jax.org>) and UniProt (<http://www.uniprot.org>), and integrated them into one in-house database. This annotation data can be used to identify terms/pathways that are enriched for genes that show evidence of association. Given a ranked list of genes, from for example an expression or SNP study, terms are assessed based on their tendency of having high-ranked genes. For this purpose one may use a hypergeometric test that require a preselection of a set of genes of interest, or perform a Wilcoxon-rank-sum test that test whether the gene ranks belonging to a term are sampled from a non-random (non-uniform) distribution. As many terms are evaluated multiple testing needs to be performed. Recent development of techniques inspired by such conventional pathway analysis indicate that this is a fruitful approach to analyse rare variants, as their combined effect is jointly considered in the context of networks.<sup>14</sup>

Given a system-wide network, such as the human interactome, one may wish to identify neighborhoods within the network that have a higher density of high-ranked genes without having to predefine subnetworks. In paper III one of the sub-analysis involved searching for subnetworks enriched for knock-out mouse genes affecting testicular development. To assess the empirical proteome-wide significance a null-distribution was created by sampling random sets of genes and estimating their enrichment. Such sampling has the benefit of retaining the topological properties of the interaction network.

## CNV locus detection

In paper II and IV identification of CNV loci was performed by the circular binary segmentation (CBS),<sup>43</sup> and algorithms that utilize hidden Markov models (HMM).<sup>9,11</sup>

HMMs are related to the EM algorithm as state assignment for a given observation needs to be determined, as well as parameters of the distribution belonging to a given state, but in HMMs the state assignments are related through a Markov process rather than independent of each other. Therefore the state assignments are specified by transition probabilities between states. The parameters of the distribution governing the observations belonging to a given state is referred to as

emission probabilities. To infer the maximum likelihood estimate of the parameters of a HMM given a set of output sequences, a type of generalized EM-algorithm called the Baum-Welch algorithm can be used. Once the transition and emission probabilities parameters are known, the Viterbi algorithm can be used to infer the most probable states of a given observed sequence.

CBS is a somewhat less sophisticated heuristic which recursively identifies breakpoints between segments of differing mean copy number values. A benefit of segmentation algorithms such as CBS is that they do not assume discrete copy number states. This is of specific use for samples from tumor biopsies, as they contain a mixture of cell populations, both with regard to cell-type and with regard to aneuploidy which typically vary within a cancer cell population.

**Part II**

**Papers**



---

## Chapter 5

# Expression profiling of testicular carcinoma *in situ*

---

### 5.1 Abstract

Testicular germ cell cancers in young adult men derive from a precursor lesion called carcinoma *in situ* (CIS) of the testis. CIS cells were suggested to arise from primordial germ cells or gonocytes. However, direct studies on purified samples of CIS cells are lacking. To overcome this problem, we performed laser microdissection of CIS cells. Highly enriched cell populations were obtained and subjected to gene expression analysis. The expression profile of CIS cells was compared with microdissected gonocytes, oogonia, and cultured embryonic stem cells with and without genomic aberrations. Three samples of each tissue type were used for the analyses. Unique expression patterns for these developmentally very related cell types revealed that CIS cells were very similar to gonocytes because only five genes distinguished these two cell types. We did not find indications that CIS was derived from a meiotic cell, and the similarity to embryonic stem cells was modest compared with gonocytes. Thus, we provide new evidence that the molecular phenotype of CIS cells is similar to that of gonocytes. Our data are in line with the idea that CIS cells may be gonocytes that survived in the postnatal testis. We speculate that disturbed development of somatic cells in the fetal testis may play a role in allowing undifferentiated cells to survive in the postnatal testes. The further development of CIS into invasive germ cell tumors may depend on signals from their postpubertal niche of somatic cells, including hormones and growth factors from Leydig and Sertoli cells.



## 5.2 Manuscript

### Analysis of gene expression profiles of microdissected cell populations indicates that testicular carcinoma *in situ* is an arrested gonocyte

Si Brask Sonne<sup>1,#</sup>, Kristian Almstrup<sup>1</sup>, Marlene D. Dalgaard<sup>1</sup>, Agnieszka S. Juncker<sup>2</sup>, Daniel Edsgård<sup>2</sup>, Ludmila Ruban<sup>3</sup>, Neil J. Harrison<sup>4</sup>, Christian Schwager<sup>5</sup>, Amir Abdollahi<sup>5</sup>, Peter E. Huber<sup>5</sup>, Søren Brunak<sup>2</sup>, Lise Mette Gjerdrum<sup>6</sup>, Harry D. Moore<sup>4</sup>, Peter W. Andrews<sup>4</sup>, Niels E. Skakkebæk<sup>1</sup>, Ewa Rajpert-De Meyts<sup>1</sup>, Henrik Leffers<sup>1</sup>

<sup>1</sup> Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.

<sup>2</sup> Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

<sup>3</sup> Department of Biochemical Engineering, University College London, London, UK

<sup>4</sup> Centre for Stem Cell Biology, University of Sheffield, Sheffield, UK

<sup>5</sup> Department of Radiation Oncology, German Cancer Research Center, Heidelberg, Germany

<sup>6</sup> Department of Pathology, Rigshospitalet, Copenhagen, Denmark

# Correspondence should be addressed to si@bsonne.dk

### Introduction

Testicular germ cell cancer is the most common malignant disease among young adult men in Europe, affecting up to 1% of all men.<sup>1</sup> All testicular germ cell tumors of young adult men derive from carcinoma *in situ* (CIS). The CIS cells are believed to arise from fetal germ cells and reside dormant in the testis until they start proliferating after puberty and eventually develop into an overt tumor.<sup>2</sup> Overt testicular germ cell tumors can be divided into two major classes: the seminomas, which retain a CIS-like phenotype and germ cell features; and the more pluripotent embryonic stem cell (ESC)-like nonseminomas, which consist of tumors resembling embryonic tissues (e.g., embryonal carcinoma and teratoma) as well as extraembryonic tissues (e.g., choriocarcinoma and yolk sac tumor).

Testicular germ cell tumors are part of the testicular dysgenesis syndrome,<sup>3</sup> a group of disorders believed to arise as a result of disturbed development of the somatic cells in the gonad, probably due to an imbalanced hormonal environment of the fetus.<sup>4</sup> The exact trigger for the neoplastic transformation is unknown, but it is probably initiated at the stage of primordial germ cells or gonocytes. This assumption is based on the morphology of CIS<sup>5</sup> and overlap in expression of markers in CIS, primordial germ cells, and gonocytes, but not in infantile spermatogonia and adult germ cells, including several embryonic pluripotency genes.<sup>6</sup> In accordance,

our recent study showed a striking resemblance between the gene expression profiles of CIS and ESCs because up to 34% of the identified CIS genes were previously reported in ESCs.<sup>7</sup> Further, when ESCs are cultured for a prolonged time, gain of chromosome arms 17q and 12p is repeatedly observed.<sup>8</sup> Interestingly, the same chromosomal regions are implicated in the progression of CIS to invasiveness, emphasizing the resemblance between CIS and ESCs.<sup>9,10</sup>

When the primordial germ cells migrate through the hindgut toward the gonadal ridge, they remain sexually bipotent. After an initial proliferation in the gonadal ridge, the female germ cells, oogonia, enter meiosis, while male germ cells, gonocytes, continue to proliferate until their differentiation to the quiescent spermatogonia. One possible explanation for the development of CIS could be that an insufficient virilization of somatic cells surrounding the germ cells could lead to a more female-like differentiation and perhaps a premature initiation of meiosis.<sup>11</sup>

Due to the cellularity of the testis, where CIS cells maximally constitute about 5% of the cells, it is difficult to make a satisfactory expression profile of CIS. Previous studies of global gene expression in CIS cells have analyzed testis tissues containing increasing proportions of CIS cells<sup>7</sup> or simply compared testis tissue with CIS to normal testis tissue.<sup>12,13</sup> Although giving useful results, these approaches are limited by a considerable background noise from other cell types in the testis.

We have addressed this issue by developing a fast and specific staining procedure for CIS and fetal germ cells,<sup>14</sup> allowing laser microdissection and RNA isolation from relatively pure cell populations. This resulted in RNA of a quality sufficient to perform two rounds of amplification, producing microgram amounts of RNA, which allowed microarray analysis.

In this study, we aimed at elucidating the origin of CIS cells based on comparative gene expression profiling. For this purpose, we compared the gene expression profiles of microdissected CIS cells, gonocytes, and oogonia and cultured ESCs with and without genomic aberrations. To correct for contamination with RNA from Sertoli cells, in which gonocytes and CIS cells are embedded, we also microdissected Sertoli cells from tubules with CIS and included these data in the analysis.

## Material and Methods

### Tissue samples and ESC lines

The Regional Committee for Medical Research Ethics in Denmark approved the use of adult testicular samples, and collection of human fetal gonads in the United Kingdom was done in agreement with the Polkinghorne guidelines, following ethical approval and informed consent of women who underwent elective abortions at 10 to 12 wk of pregnancy. Adult testicular tissues containing CIS were residual tissues from orchidectomies collected at the Department of Pathology at Rigshospitalet after diagnosis of testicular cancer. One of the normal testis RNA samples was from apparently normal tissue adjacent to a tumor, and the other two were commercial samples (Applied Biosystems/Ambion and Clontech). Fetal gonads were collected in Sheffield from abortion material; the gestational age was estimated by ultrasound scanning and measurements of hand size. Cultures of the human ESC

lines (H7 and Shef5) were maintained in Sheffield under the direction of two of the authors (P.W.A. and H.D.M.). H7 cells with or without genomic aberrations were sorted using a fluorescence-activated cell sorter and only cells expressing the pluripotency marker SSEA3 (undifferentiated cells) were analyzed (see Table 5.1 for a more thorough description of the samples).

### Preparation of cryosections for microdissection

Adult testicular tissue and fetal gonads were embedded in optimum cutting temperature compound (Sakura Fintek Europe) and snap frozen at  $-80^{\circ}\text{C}$  in isopentan. Sections of  $10\ \mu\text{m}$  (fetal tissue) and  $20\ \mu\text{m}$  (adult tissue) were cut on a Shandon SME Cryotome (Life Sciences International Europe Ltd.), collected on nuclease- and nucleic acid-free membrane slides (Molecular Machines & Industries), immediately fixed in 75% RNase-free ethanol for 10 min, and stored in absolute ethanol at  $-80^{\circ}\text{C}$ . Serial sections of fetal gonads and testicular tissues containing CIS were analyzed by immunohistochemistry (Figure 5.1A) for AP-2g to identify gonocytes, oogonia, and CIS;<sup>15</sup> fetal antigen-1 (FA-1) to identify Leydig cells;<sup>16</sup> and AMH<sup>17</sup> and MIC-2<sup>18</sup> to identify fetal and adult Sertoli cells, respectively. An additional serial section was stained for alkaline phosphatase activity,<sup>14</sup> which is only detectable in CIS and fetal germ cells.

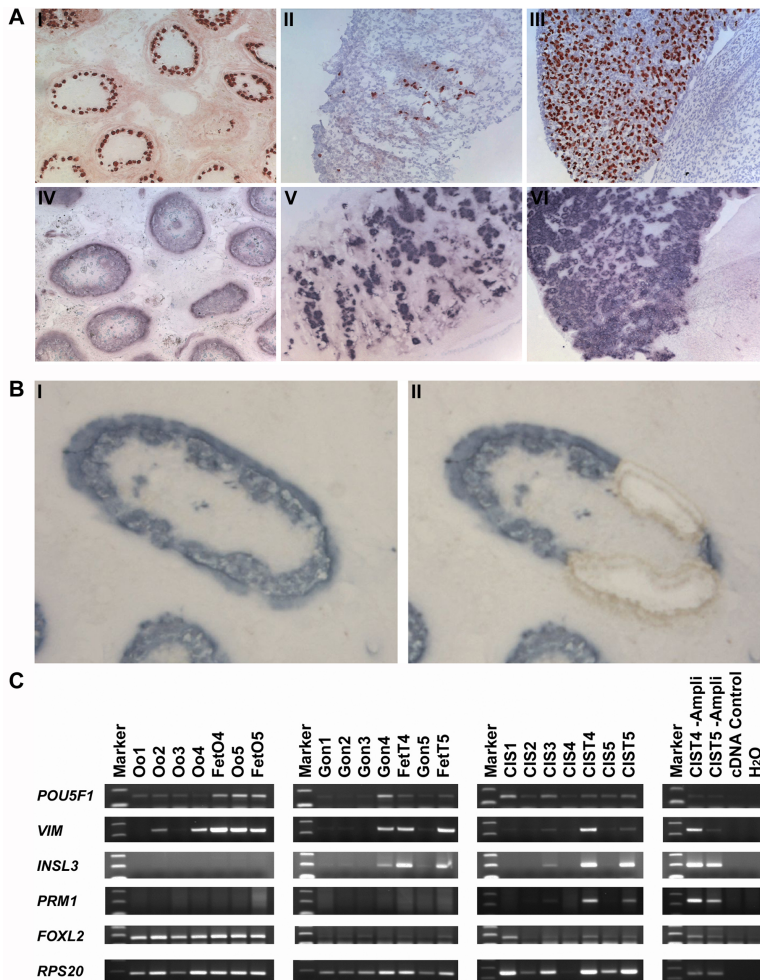
Microdissection and RNA amplification. Before microdissection, slides were transferred to room temperature and stained with nitroblue tetrazolium (NBT)-5-bromo-4-chloro-3-indolyl phosphate (BCIP) by direct histochemistry as previously described.<sup>14</sup> The cells were microdissected within 2 h at room temperature using the MMI CellCut or SmartCut system (Olympus/Molecular Machines & Industries; Figure 5.1B). Only CIS tubules with a classic appearance with CIS cells along the edge of tubules and stained areas that resembled fetal germ cells were excised to avoid CIS cells at a more advanced stage or unspecifically stained areas.

RNA was purified using the Ambion RNAqueous Micro Kit (Applied Biosystems/Ambion). The RNA quality was tested with the Bioanalyzer Picokit (Agilent Technologies), and samples were amplified in two rounds using the MessageAmp II aRNA Amplification Kit (Applied Biosystems/Ambion).

### Microarray analysis

The following samples were analyzed: three ESC samples, three microdissected oogonia samples, three microdissected gonocyte samples, three microdissected CIS samples (CIS), and three microdissected Sertoli cell samples from tubules containing CIS. In addition, three samples of testis tissue containing CIS (CIST) from the same patients as the microdissected CIS and three normal testis samples were included (Table 5.1). All samples underwent two rounds of amplification as described above.

For microarray analysis, we used Agilent Whole Human Genome Microarray 4 $\times$ 44K chips (design no. 014850, Agilent Technologies). Hybridization and scanning of one-color arrays were done as described by the manufacturer (Agilent Technologies) and analyzed using the Agilent Feature extraction software (version 9.1.3.1).



**Figure 5.1** – Verification of microdissection. **A**, serial sections of a testis with CIS (I and IV), fetal testis (II and V), and fetal ovary (III and VI). Top, immunohistochemical staining for AP-2 $\gamma$ ; bottom, alkaline phosphatase expression visualized by NBT-BCIP staining. **B**, frozen section with CIS stained with NBT-BCIP before (I) and after (II) laser microdissection. **C**, RT-PCR of representative genes. Tissues 1 to 3 are the same samples used in the microarray analysis; tissues 4 and 5 are amplified RNA from microdissected and total tissues; in the left panel, CIS 4 and 5 are unamplified whole testis samples from the same patients as in the previous panel (see Table 5.1 for a more thorough description). Oo, microdissected oogonia; FetO, fetal ovary; Gon, microdissected gonocytes; FetT, fetal testis; -Ampli, not amplified.

The lowess normalized, `gProcessedSignal` from each array was loaded into the `marray` and `limma R/BioC` package, normalized between arrays using a quantile normalization procedure, and log transformed. Normalized data were then imported into TIGRs MeV v4.0<sup>19</sup> for subsequent statistical analysis using the significance analysis of microarrays (SAM<sup>20</sup>) with standard settings. Partition clustering of selected gene lists was made by the Partitioning Around Medoids clustering algorithm (R library “`cluster`”), and Correspondence Analysis (R library “`MASS`”<sup>21</sup>) was done to elucidate the correspondence between profiles of selected genes across a set of cell types.

### Reverse transcription-PCR, immunohistochemistry, and *in situ* hybridization

cDNA synthesis was made with 50 ng/AL random hexamer primers. Reverse transcription-PCR (RT-PCR) was done using gene-specific primers placed just upstream of the polyA site. Primer sequences, cycle numbers, and annealing temperatures are summarized in Supplementary Table S1. RPS20 was used as a positive control for cDNA synthesis and PCR efficiency. Representative bands for each primer set were excised from the gels and sequenced.

Serial sections of frozen tissues (10 Am) used for microdissection were fixed in phosphate buffered formalin (4% w/v, pH 7; VWR, Bie & Berntsen) for 10 min. Immunohistochemistry was done as previously described<sup>17</sup> with the following antibodies: AP-2g (Santa Cruz Biotechnology), FA-1 (provided by Charlotte Harken Jensen, Odense University, Odense, Denmark), AMH (provided by Richard L. Cate, Biogen, Cambridge, MA), MIC-2 (clone 12E7, Dako), and MYCL1 [L-Myc (C-20), Santa Cruz Biotechnology]. The antibodies were diluted 1:50, 1:200, 1:150, 1:50, and 1:100, respectively.

*In situ* hybridization was done as previously described.<sup>22</sup> The DNA template for the THC2340734 RNA probe was amplified using nested primers: 1st PCR, GC-CAACAAGAAGGACATCATGA and GTATGGGAATGGATGGTGTGTG; 2nd PCR, AATTAACCCTCACTAAAGGGACATCATTGACCCT and TAATACGACTCAC-TATAGGGTGTGTTCCCTGT (boldface indicates T3 and T7 promoters).

## Results

### Purity of microdissected CIS and fetal germ cells

Although it was impossible to completely avoid contamination with surrounding somatic cells, we obtained enriched cell populations that, based on visual inspection, contained up to 60% oogonia, 80% gonocytes, and 80% CIS cells (Figure 5.1; Table 5.1). Before we proceeded to microarray analysis, we tested the enrichment of RNA from CIS and fetal germ cells by RT-PCR analysis of genes with cell type-specific expression (Figure 5.1C). The selected genes were *POU5F1* (*OCT3/4*), expressed in CIS, gonocytes, and oogonia;<sup>23</sup> *VIM*, expressed in Sertoli cells, endothelial cells, Leydig cells,<sup>24</sup> and granulosa cells;<sup>25</sup> *INSL3*, specific for Leydig cells;<sup>26</sup> *FOXL2*, expressed in granulosa cells and undifferentiated Sertoli cells in fetal testes and adult testes with signs of testicular dysgenesis syndrome;<sup>27</sup> and

*PRM1*, expressed in round and elongating spermatids.<sup>28</sup> *RPS20* was used as a cDNA synthesis and PCR control.

*POU5F1* was, as expected, present in microdissected samples of oogonia, gonocytes, and CIS, and when compared with *RPS20*, the bands were generally stronger in the microdissected samples than in the whole tissue preparations. *VIM* was present in the fetal ovary, fetal testis, and whole CIS testis preparations, but weak bands were also seen in some of the microdissected oogonia, gonocyte, and CIS samples, indicating contamination with neighboring Sertoli and granulosa cells. The Leydig cell gene *INSL3* was only detected in samples containing total testis RNA and, at a very low level, in one of the CIS and a few of the gonocyte samples. *FOXL2* was detected in all oogonia samples, reflecting the granulosa cell contamination, and was also weakly detectable in gonocyte and CIS samples, probably due to contamination with undifferentiated and partially undervirilized Sertoli cells.<sup>27</sup> *PRM1* was, as expected, expressed in all CIST samples because of the presence of normal tubules with spermatogenesis, but not in any of the microdissected or fetal samples.

Based on these results, we concluded that the microdissected cell populations were adequately enriched to proceed to microarray analysis.

### Evaluation of microarray gene expression profiles

Gene expression was analyzed using Agilent microarrays covering more than 41,000 unique human genes and transcripts; the analyzed samples are summarized in Table 5.1. The raw data have been submitted to Array Express at the European Bioinformatics Institute (accession no. E-TABM-488).

To investigate if previously described CIS markers were also identified in this data set, we performed a two-way SAM of microdissected CIS versus normal testis (Figure 5.2A). Of the 26 most significant genes [false discovery rate (FDR) = 0.1%], five genes had been identified as CIS markers in previous studies [*PDPN* (also known as M2A antigen<sup>29</sup>), *TFCP2L1*, *IL22RA1*, *UPP1*, and *MYCL1* (previously *MYCL*<sup>7,13</sup>)] and one (*ADORA2B*) had been reported in undifferentiated testicular tumors and cell lines.<sup>13</sup> By comparison, similar two-way SAM analyses (with the same FDR of 0.1%) on CIST versus normal testis and CIST + CIS versus normal testis resulted in zero significant genes specific for CIS.

*MYCL1* seemed to be highly expressed in CIS cells, and immunohistochemistry confirmed its presence in CIS cells (Figure 5.2B). The heatmap of genes highly expressed in CIS included five tentative human consensus (THC) sequences that seemed to be largely restricted to CIS cells (Figure 5.2A). This was confirmed by RT-PCR for three of them, whereas one (THC2341283) did not give any bands and one (THC2378933) resulted in multiple bands (not shown). *In situ* hybridization with a probe for THC2340734 confirmed its high expression in CIS cells (Figure 5.2C). However, all three THC transcripts only included a single exon, TBD:7 and none seemed to encode open reading frames.

Among the genes very high in CIS compared with normal testis was a Sertoli cell marker, *KRT16* (immunohistochemistry not shown), indicating a severe contamination of CIS samples with the neighboring Sertoli cells. However, a blast

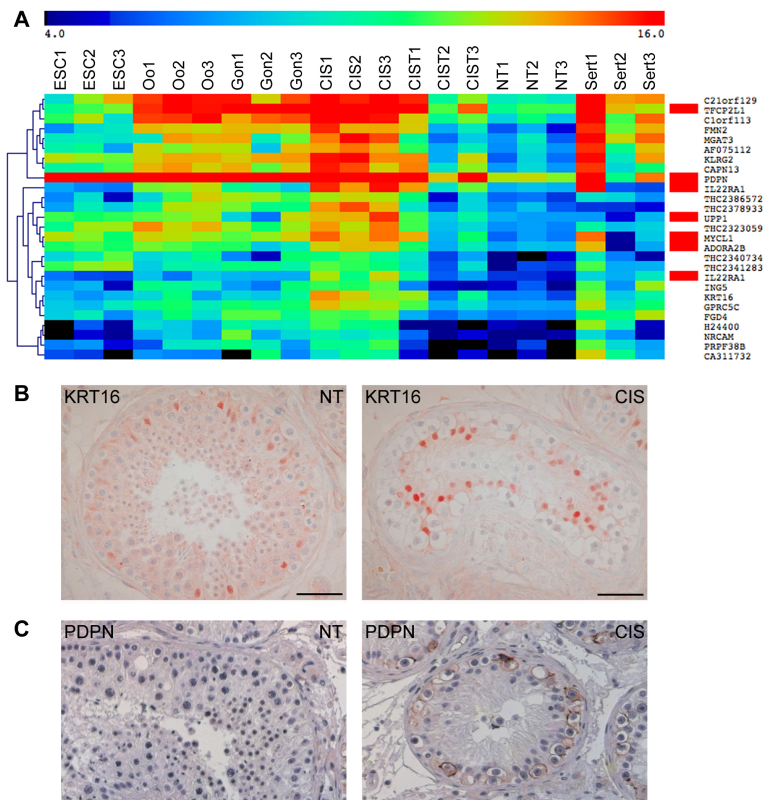
search at ENSEMBL revealed that the *KRT16* oligo was only 95% identical to *KRT16* (three mismatches), but had a 100% match to two other transcripts that seemed to be nonfunctional *KRT16* pseudogenes. To avoid similar mistakes for other genes, we checked all the oligo probes shown in the heatmaps and found that only 77% of the oligo probes for annotated genes were unique for their intended targets (indicated by asterisks). However, because the results in this study are not based on expression of single genes but on the overall profile, the lack of specificity for a subset of the oligos does not affect the results.

The CIS samples clearly included material from Sertoli cells, and therefore we determined the expression profiles of microdissected Sertoli cells from CIS tubules and tested various methods for subtracting Sertoli genes from the CIS data. However, irrespective of the method, this also subtracted a large number of genes expressed in both Sertoli and CIS cells. Instead, we decided to include the expression profile of microdissected Sertoli cells in the heatmap to facilitate a visual comparison (Figure 5.2A).

### CIS cells are very similar to gonocytes

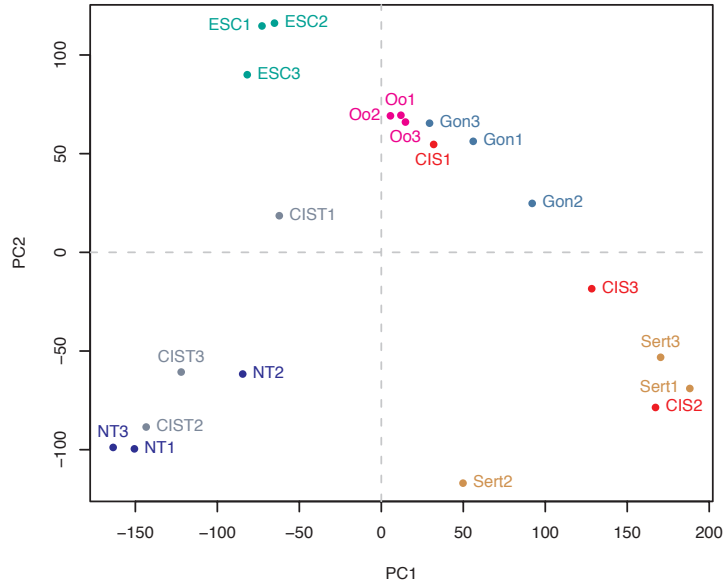
A principal component analysis of the entire data set (Figure 5.3) showed that the expression profiles generally grouped by cell type. Normal testis, ESC, and Sertoli were the most distant groups, and oogonia and gonocytes grouped closely together. The gene expression profiles for CIS were clearly distinct from the CIST samples, whose profiles resembled normal testis whereas the microdissected CIS samples grouped with fetal samples and Sertoli cells. This confirmed that the microdissection was successful in eliminating normal germ cells from the CIS gene expression profile. Interestingly, the content of CIS in CIST samples was reflected in the plot; CIST2, which contained only 15% CIS tubules, was closer related to normal testis, whereas CIST1, containing 95% CIS, was much closer related to the fetal samples and microdissected CIS. The overlap between CIS and Sertoli cells was probably due to different levels of contamination with Sertoli cells. Sample clustering of the 500 most differentially expressed genes from a six-way SAM (FDR = 0%;  $\delta = 0.36$ ; Supplementary Figure S1) gave similar results: The gene expression profile of CIS cells was very similar to that of gonocytes and also related to the oogonial expression profile, whereas CIST and normal testis clustered together, and ESCs were more distantly related to these tissues (Supplementary Figure S1). The similarity between CIS and gonocytes was further supported by analysis of uncorrelated shrunken centroids.<sup>30</sup> Leaving Sertoli samples out of the uncorrelated shrunken centroid analysis, all the CIS samples were classified as gonocytes, and all CIST samples as normal testis. Thus, based on gene expression profiles, CIS cells are most similar to gonocytes.

We made additional analyses to characterize the data set: Supplementary Figure S1 shows a heatmap of the 500 most differentially expressed genes from a six-way SAM; Supplementary Figure S2 shows the most frequent profiles among these genes; and Supplementary Figure S3 shows a heatmap of genes specific for the respective cell types characterized in this study.



**Figure 5.2** – Verification of microarray data, refining CIS genes. **A**, two-way SAM on CIS versus normal testis ( $\delta = 0.84$ ; FDR = 0.1%) showing genes highly expressed in CIS; note that IL22RA1 is detected with two different oligo probes. Genes marked with red were previously described in CIS or undifferentiated testicular tumors and cell lines. The annotated genes with confirmed probe specificity are marked by asterisks. The color key at the top shows the relative expression levels from 4 to 16. A more extensive gene list (FDR <1%) can be found in Supplementary Table S2. **B**, expression of MYCL1 in CIS testis. **C**, expression of the THC2340734 transcript in CIS testis. Arrows, CIS cells; arrowheads, Sertoli cells. Note that MYCL1 is also expressed in interstitial Leydig cells. NT, normal testis; Sert, microdissected Sertoli cells next to CIS (the rest of abbreviations are as in the legend to Figure 5.1).





**Figure 5.3** – Principal component analysis. All the samples included in the microarray data set were subjected to a principal component analysis to determine their spatial relationship to each other.

### Identification of gene clusters possibly involved in the origin of CIS

Figure 5.4 shows biologically interesting clusters that may provide additional knowledge on how CIS arise.

When CIS and gonocytes were compared, only five transcripts came out as differentially expressed (FDR <1%). Two genes were up-regulated in CIS: *DEFB119*, encoding an antimicrobial peptide, regulated by androgens and specifically expressed in the testes;<sup>31</sup> and *NMNAT2* (nicotinamide mononucleotide adenylyl-transferase-2), a central enzyme of the NAD biosynthetic pathway.<sup>32</sup> Three genes were down-regulated in CIS versus gonocytes: *PTPRZ1*, a protein tyrosine phosphatase receptor, which has been described in several cancer types;<sup>33</sup> the predicted cancer associated gene *ASXL3*, a human homologue of the *Drosophila* additional sex combs (asx) gene;<sup>34</sup> and one unannotated gene (AF318333).

An interesting finding was a large cluster of genes that were expressed at a lower level in CIS compared with the other germ cell types. This has not previously been described because it requires pure CIS preparations. These genes may give important clues to the mechanisms of the neoplastic transformation to CIS because their down-regulation may be linked to this process.

To further study the relationships between CIS, fetal gonocytes, and fetal oogonia, we performed a three-way SAM and selected the 62 most significant genes

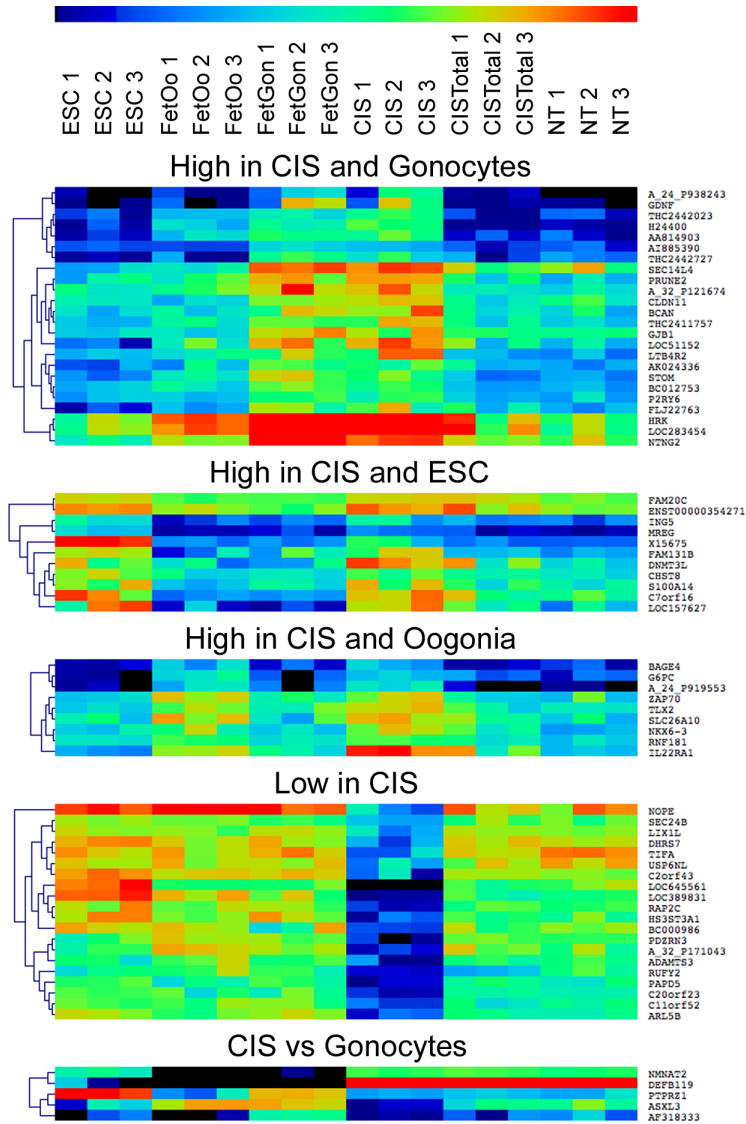
for correspondence analysis and partition clustering (Figure 5.5). Comparing the three profiles of genes upregulated in CIS, gonocytes, and oogonia, it was obvious that the genes representing the gonocyte and oogonia profiles were expressed at a lower level in all other cell types, whereas the genes in the CIS profile were also highly expressed in normal adult testis and in Sertoli cells. A relatively large cluster contained genes up-regulated in CIS and gonocytes and down-regulated in oogonia. Not surprisingly, among the 13 genes in this cluster, 6 were located on the Y-chromosome, consistent with previous observations<sup>35</sup> that CIS arises only in individuals with some Y chromosome material present.

## Discussion

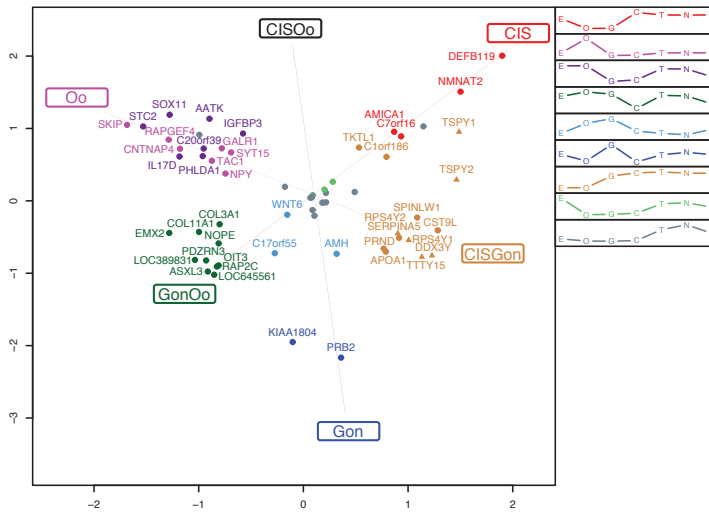
With this study, we, for the first time, performed gene expression analysis on isolated CIS and fetal germ cells, allowing direct comparison of gene expression profiles of CIS cells, gonocytes, and oogonia. The results were clear: No matter how the data were analyzed, CIS cells always grouped with gonocytes. This study supports the proposed fetal origin of CIS and provides the basis for a more detailed understanding of CIS.

Numerous previous studies of single genes clearly showed that CIS cells in many aspects resemble gonocytes<sup>6,15,29,36-41</sup> and that both CIS and gonocytes express a number of genes normally only seen in ESCs.<sup>23,39,40</sup> This was confirmed in earlier microarray studies wherein gene expression in whole testes with CIS was compared with normal testes.<sup>7,12,13</sup> In the study by Almstrup and colleagues,<sup>7</sup> we analyzed three testis samples with increasing amounts of CIS tubules and sorted the data according to the percentage of CIS cells. Among the 100 genes most highly expressed in CIS compared with normal testis, 34 genes were also reported in ESCs. However, according to the present data set, CIS cells are much more similar to gonocytes than to ESCs. Although we cannot exclude that the *in vitro* adaptation of ESCs and contamination with Sertoli cells may add to the difference between CIS and ESCs, we believe that the previously reported similarity between these cells was caused by the study design,<sup>7,12,13</sup> which led to a subtraction of genes expressed in both normal testis and CIS. This resulted in an overrepresentation of pluripotency genes and, thus, a more pronounced similarity to ESCs.

By isolating the different cell types, we found that gonocytes and CIS cells were very similar and that both were closely related to oogonia, which have only recently diverged from the gonocytes in the sex differentiation. However, according to the correspondence analysis on genes differentially expressed between CIS, oogonia, and gonocytes (Figure 5.5), no genes distinguished CIS and oogonia from gonocytes, indicating that CIS do not originate from cells with oogonia characteristics, whereas several genes were highly expressed in CIS and gonocytes, but not in oogonia. Only five genes significantly distinguished CIS from gonocytes (Figure 5.4), and interestingly, among these five genes, two cancer-associated genes had low expression levels in CIS. Moreover, among the genes highly expressed in CIS, we found no overrepresentation of oncogenes, and among the genes with low expression in CIS, we found no overrepresentation of tumor suppressor genes. This indicates that CIS is not a malignant cancer cell in a classic sense but rather an arrested gonocyte.



**Figure 5.4** – Biologically interesting clusters. A two-way SAM of interesting cell combinations. CIST was excluded in all analyses. From the top: CIS + gonocytes versus others ( $\delta = 1.21$ ; FDR = 0.314%), CIS + ESC versus others ( $\delta = 1.24$ ; FDR = 0.964%), CIS + oogonia versus others ( $\delta = 1.33$ ; FDR = 0%), CIS versus others ( $\delta = 3.27$ ; FDR = 0%), and CIS versus gonocytes ( $\delta = 3.90$ ; FDR = 0%). Annotated genes with confirmed probe specificity are marked by asterisks. The color key at the top shows the relative expression levels from 4 to 16. Abbreviations are as in the legends to previous figures. Only the most differentially expressed genes are shown in Figure 5.4; more extensive analyses are available in Supplementary Table S3.



**Figure 5.5** – Correspondence analysis of genes characterizing CIS, gonocytes, and oogonia. Correspondence analysis for the 62 most significant genes from a three-way SAM on the classes CIS, gonocytes, and oogonia. The 62 genes were clustered into nine distinct gene groups (visualized by coloring), and the corresponding mean cluster profiles are displayed as legend. The positions of the cluster names (CIS, CISOo, etc.) represent the ideal profiles for genes expressed only in the particular cell types. 4, gene located on the Y chromosome. E, ESCs; O, oogonia; G, gonocytes; C, carcinoma in situ ; T, testis tissue containing CIS; N, normal testis; S, Sertoli cells next to CIS; CISOo, CIS and oogonia; GonOo, gonocytes and oogonia; CISGon, CIS and gonocytes.

There have been other proposals for the origin of CIS. Clark<sup>42</sup> suggested that CIS may originate from a multipotent spermatogonial stem cell. However, human spermatogonia do not express multipotency genes characteristic of CIS, and although multipotent stem cells have been derived *in vitro* from adult human testis,<sup>43</sup> the expression profile of these cells does not correspond to the profile of CIS cells. Most noteworthy, the CIS markers *POU5F1*, *NANOG*, and *CDH1* were only expressed at a low level, and *KLF4* and *STAT3*, which are expressed at a very low level in CIS according to the present data set, were highly expressed in the spermatogonial cell population and adult germ line stem cells. Thus, this origin is rather unlikely.

Some CIS cells are hypertriploid,<sup>44,45</sup> and this feature was suggested as an argument for their origin from spermatocytes, which duplicate their genome in preparation for meiosis.<sup>46</sup> Alternatively, CIS cells could originate from fetal gonocytes that, in analogy to female oogonia, attempted to enter meiosis in fetal life because of insufficiently virilized microenvironment in dysgenetic testis.<sup>11</sup> The polyploidization followed by selective gene losses and gains have been proposed by several studies as the earliest abnormality in CIS cells<sup>10,47,48</sup> and has been detected even in pre-CIS, an abnormal germ cell in severely dysgenetic subjects with disorders of

sex differentiation.<sup>49</sup> However, the hypertriploidy of CIS cells does not need to be related to meiosis because most other cancers are also aneuploid. Accordingly, we found no meiosis-related genes among the genes specific for CIS. Moreover, when we selected genes highly expressed in CIS and normal adult testis but with low expression in gonocytes (Supplementary Table S4), we did not find any genes specific for meiosis, which strongly suggest that the cell of origin is not a meiotic cell.

### **Pitfalls in the microdissection technology**

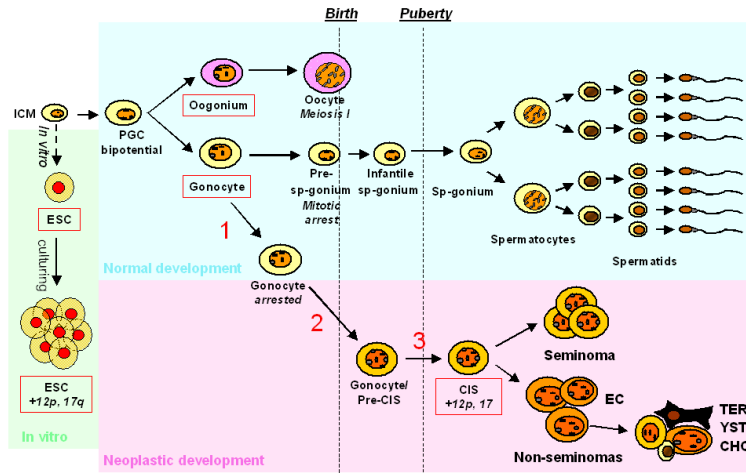
Several challenges with the microdissection technology must be taken into consideration. Especially, it is necessary to verify the enrichment of RNA from target cells by RT-PCR before expression profiling. The preliminary verification in this study was encouraging, but also showed that it is impossible to obtain completely pure cell populations by laser microdissection of tissue sections because inclusion of material from neighboring cells is unavoidable and cellular material from cells cut open by the microtome will inevitably leak to the surrounding tissue. Nevertheless, both RT-PCR and subsequent microarray analysis showed a substantial enrichment of RNA from target cells, and expression profiling clearly showed different profiles for different cell types. We also observed specificity problems with the Agilent oligos. Many probes either matched multiple distinct transcripts or recognized opposite strands or introns in their designated targets. This underlines the importance of checking that the oligos match their targets, and further emphasizes that microarray results should always be verified. However, in this study, we do not focus on expression of single genes; instead, we compare global expression profiles, which leads to quite robust results that are not affected by uncertainty about the identity of a few genes.

### **Origin of CIS**

Although this study confirms previous candidate gene-based studies and clearly shows that CIS cells are very similar to gonocytes, it still remains to be determined why these gonocytelike cells do not differentiate to spermatogonia but persist in postnatal testes. We know many risk factors for testis cancer, and virtually all are related to a poor embryonic and fetal development of the testes. Animal studies suggest that this may be related to reduced testosterone levels causing a maldevelopment of somatic cells of the testis.<sup>50</sup> The poor function of the somatic cells affects the germ cells, which fail to differentiate properly without the appropriate paracrine signals from Sertoli and peritubular myoid cells.<sup>4</sup> Thus, we suggest that CIS originates from gonocytes that failed to differentiate to pre-spermatogonia due to the undermasculinized somatic cells that (a) fail to stimulate the germ cells sufficiently and (b) constitute a microenvironment that allow fetal gonocytes to survive in the postnatal testes (Figure 5.6). Further investigation of the somatic compartment and its role in the progression from gonocyte to pre-spermatogonium will aid in the understanding of CIS.

### **Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.



**Figure 5.6** – Schematic illustration of the hypothesized origin of CIS from gonocytes. Schematic illustration of the normal male germ cell development and possible transformation to CIS. ESCs are derived from the inner cell mass of a blastocyst. Prolonged culturing often leads to an accumulation of chromosomal aberrations, especially gain of material from chromosomes 12p and 17q. During early development, primordial germ cells migrate to the gonadal ridge and develop along the female (oogonia; top) or the male (gonocytes; bottom) germ cell lineage. In the male, the gonocytes become embedded in Sertoli cells, creating testicular cords. During the period from third trimester to 3 mo postnatally, the gonocytes migrate to the periphery of the tubules and differentiate to pre-spermatogonia. After puberty, the spermatogonia proliferate and start spermatogenesis. CIS cells are proposed to arise when gonocytes fail to differentiate to pre-spermatogonia (1) and fail to undergo apoptosis (2). These gonocytes or pre-CIS cells lie dormant in the testis through infancy, while genomic aberrations may occur (3); at puberty, when testosterone levels increase, they start to proliferate and genomic aberrations accumulate, especially of chromosome 12p and 17, eventually resulting in the formation of an overt tumor. ICM, inner cell mass; PGC, primordial germ cells; Sp-gonium, spermatogonium; EC, embryonal carcinoma; TER, teratoma; YST, yolk sac tumor; CHC, choriocarcinoma.

### Acknowledgments

Received 12/2/08; revised 3/28/09; accepted 4/16/09; published Online 6/2/09.

Grant support: Danish Cancer Society, The Villum Kann Rasmussen Foundation, Svend Andersen's Foundation, and Kirsten and Freddy Johansen's Foundation. The work by Peter Andrews and Neil J. Harrison was funded by the Medical Research Council and Yorkshire Cancer Research.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Sabina Soultanova, John Nielsen, Anne Jørgensen, Brian Vendelbo Hansen, Betina Nielsen, Muaber Zejnuli, Thomas Regiert, Heiko Drzonek, and Ute Wirkner for skillfull technical assistance.

### 5.3 Supplementary Material

Available online from the journal website.<sup>1</sup>

---

<sup>1</sup><http://cancerres.aacrjournals.org/content/69/12/5241/suppl/DC1>

Sample ID	Age	Description	Estimated purity (%)	Ampli	RT-PCR	Array
ESC1	—	H7 abnormal subline, SSEA3-positive	100	Yes	—	Yes
ESC2	—	H7 normal subline, SSEA3-positive	100	Yes	—	Yes
ESC3	—	Shef5, not fluorescence-activated cell sorted	100	Yes	—	Yes
Oo1	12–13 wg	Microdissected fetal oogonia	60	Yes	Yes	Yes
Oo2	11–12 wg	Microdissected fetal oogonia	60	Yes	Yes	Yes
Oo3	10–11 wg	Microdissected fetal oogonia	60	Yes	Yes	Yes
Oo4	11–12 wg	Microdissected fetal oogonia	60	Yes	Yes	—
FetO4	11–12 wg	Total fetal ovary, same tissue as Oo4	—	Yes	Yes	—
Oo5	11–12 wg	Microdissected fetal oogonia	60	Yes	Yes	—
FetO5	11–12 wg	Total fetal ovary, same tissue as Oo5	—	Yes	Yes	—
Gon1	10–11 wg	Microdissected fetal gonocytes	80	Yes	Yes	Yes
Gon2	10–11 wg	Microdissected fetal gonocytes	80	Yes	Yes	Yes
Gon3	11–12 wg	Microdissected fetal gonocytes	80	Yes	Yes	Yes
Gon4	10–11 wg	Microdissected fetal gonocytes	80	Yes	Yes	—
FetT4	10–11 wg	Total fetal testis, same tissue as Gon4	—	Yes	Yes	—
Gon5	12–13 wg	Microdissected fetal gonocytes	80	Yes	Yes	—
FetT5	12–13 wg	Total fetal testis, same tissue as Gon5	—	Yes	Yes	—
CIS1	37 y	Microdissected CIS from tissue with CIS + invasion (seminoma-like)	80	Yes	Yes	Yes
CIST1	37 y	Same tissue as CIS1, 95% CIS tubules	—	Yes	Yes	Yes
CIS2	30 y	Microdissected CIS from tissue with CIS next to classic seminoma	80	Yes	Yes	Yes
CIST2	30 y	Same tissue as CIS2, 15% CIS tubules	—	Yes	Yes	Yes
CIS3	26 y	Microdissected CIS from tissue with CIS next to nonseminoma, predominantly EC, immature Sertoli cells 80	Yes	Yes	Yes	—
CIST3	26 y	Same tissue as CIS3, 60% CIS tubules	—	Yes	Yes	Yes
CIS4	27 y	Microdissected CIS from tissue with CIS next to nonseminoma, predominantly EC progressing t YST, focal immature TER	80	Yes	Yes	—
CIST4	27 y	Same tissue as CIS4, 30% CIS tubules	—	Yes	Yes	—
CIST4-Ampli	27 y	Same tissue as CIST4, not amplified	—	—	Yes	—
CIS5	55 y	Microdissected CIS from tissue with CIS next to nonseminoma, predominantly EC and TER	80	Yes	Yes	—
CIST5	55 y	Same tissue as CIS5, 95% CIS tubules	—	Yes	Yes	—
CIST5-Ampli	55 y	Same tissue as CIST5, not amplified	—	—	Yes	—
NT1	—	Ambion	—	Yes	—	Yes
NT2	54 y	Areas of impaired spermatogenesis, hyalinized tubules and lymphocyte infiltration	—	Yes	—	Yes
NT3	—	Clontech	—	Yes	—	Yes
Sert1	37 y	From same tissue as CIS1	80	Yes	—	Yes
Sert2	30 y	From same tissue as CIS2	90	Yes	—	Yes
Sert3	26 y	From same tissue as CIS3	90	Yes	—	Yes

**Table 5.1** – Abbreviations: Oo, microdissected oogonia; FetO, fetal ovary; Gon, microdissected gonocytes; FetT, fetal testis; -Ampli, not amplified; NT, normal testis; Sert, microdissected Sertoli cells from CIS tubules; wg, weeks gestation; EC, embryonal carcinoma; YST, yolk sac tumor; TER, teratoma.





---

## Chapter 6

# Familial copy number variation in testicular cancer

---

### 6.1 Abstract

To search for disease related copy number variations (CNVs) in families with a high frequency of germ cell tumours (GCT) we analysed sixteen individuals from four families by array comparative genomic hybridization (aCGH) and applied an integrative systems biology algorithm that prioritises risk-associated genes among loci targeted by CNVs. The top-ranked candidate, *RLN1*, encoding a relaxin-H1 peptide, although only detected in one of the families, was selected for further investigations. Validation of the CNV at the *RLN1* locus was performed as an association study using qPCR with 106 sporadic testicular GCT patients and 200 healthy controls. Observed CNV frequencies of 1.9% among cases and 1.5% amongst controls were not significantly different (odds-ratio = 1.26,  $P = 1$ ). Immunohistochemistry for Relaxin-H1 (RLN1), Relaxin-H2 (RLN2), and their cognate receptor, RXFP1, detected one, and in some cases both, of the relaxins in Leydig cells, Sertoli cells, and a subset of neoplastic germ cells, while the receptor was present in Leydig cells and spermatids. Collectively, the findings show that a heterozygous loss at the *RLN1* locus is not a genetic factor mediating high population-wide risk for TGCT, but do not exclude a contribution of this aberration in some cases of cancer. The preliminary expression data suggest a possible role of the relaxin peptides in spermatogenesis and warrant further studies.

## 6.2 Manuscript

### Heterozygous deletion at the RLN1 locus in a family with testicular germ cell cancer identified by integrating copy number variation data with phenome and interactome information.

Daniel Edsgård<sup>1,#</sup>, Maria Scheel<sup>2</sup>, Niclas Tue Hansen<sup>1</sup>, Thomas Skøt Jensen<sup>1</sup>, Søren Brunak<sup>1</sup>, Niels E. Skakkebak<sup>2</sup>, Ramneek Gupta<sup>1</sup>, Ewa Rajpert-De Meyts<sup>2</sup>, Anne Marie Ottesen<sup>2</sup>

<sup>1</sup> Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

<sup>2</sup> Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.

# Correspondence should be addressed to edsgard@cbs.dtu.dk

### Introduction

Testicular germ cell tumour (TGCT) is the most common cancer type in males aged 15-40 years and its incidence has steadily risen in many developed countries over the last decades.<sup>1</sup> The recent increase has led to postulations of a primary role of environmental or lifestyle-related factors but there is also evidence for substantial genetic contribution to TGCT susceptibility. The hereditary factor of TGCTs is the highest among all common cancers with a relative risk of 4.5 among sons of an affected father, and around 8 among brothers.<sup>2</sup> However, large linkage analysis studies have failed to identify a single locus explaining a large fraction of the familial risk, suggesting that multiple susceptibility loci with weaker effect contribute to the disease.<sup>3</sup> More recently, two genome-wide association studies (GWAS) of TGCT identified several susceptibility loci, the strongest at *KITLG* and *SPRY4*, with similar effect-sizes for familial and sporadic cases.<sup>4,5</sup> Despite these findings, further genetic factors remain to be identified in order to explain a larger fraction of the heritability.

The two reported GWASs focused on single-nucleotide polymorphisms (SNPs). However, another significant source of genetic diversity are copy number variations (CNVs),<sup>6</sup> and recent studies have associated CNVs with several diseases, including neuroblastoma,<sup>7</sup> autism<sup>8</sup> and systemic autoimmunity.<sup>9</sup> Genome-wide association studies on large case-control cohorts have frequently been applied to common diseases,<sup>10</sup> but more rare genetic diseases are often studied in family-based designs. As family members have a substantial part of their genetic material in common and sample sizes can be small, genome-wide analysis alone may not have enough power to pinpoint risk-associated genetic variations. This is a particular problem in moderately rare or heterogeneous diseases where no single loci can be found to have a strong penetrance. To accommodate this we have recently described a method which prioritises genes among a set of candidate disease susceptibility loci,<sup>11</sup> and in this study it was for the first time applied in the context of

CNVs. The algorithm builds on the hypothesis that the aetiology of complex disorders can be described in terms of deregulated networks and that genes causing similar diseases are likely to be interacting. According to this principle the method ranks genes based on their protein-network proximity to previously disease-associated genes, weighed by the phenotypic similarity to the disease under study.

To find CNVs implicated in the development of TGCT, we applied this novel algorithm to an array comparative genomic hybridization (aCGH) dataset of four families with multiple cases of germ cell tumours. We identified a candidate CNV at *RLN1* in one of the families, and validated the aberrant locus in a series of patients with sporadic TGCT and healthy controls. In addition, we performed a preliminary investigation of the immuno-expression of this gene and its receptor in the human testis with and without the TGCT precursor, carcinoma *in situ* (CIS).

## Material and Methods

### Patients and tissue samples

In total 16 individuals from four cancer-prone families, each with at least two members afflicted by TGCT (Supplementary Table 1), were assayed by aCGH. Risk-association of the CNV in the candidate locus obtained from the aCGH analysis was assessed by an association study on a case-control cohort of 106 patients previously diagnosed with sporadic TGCT, and 200 healthy male military conscripts of 18 – 21 years of age. The samples for the validation cohort were selected from DNA biobanks at Rigshospitalet, Copenhagen, Denmark. All patients and controls were ethnically Caucasian, more than 95% were Danish. None of the military conscripts had a history of neoplastic disease or cryptorchidism. In addition, eleven patients from six families with germ cell cancer and six family members without any diagnosed malignant disease were investigated in the validation stage.

A protein localization study was conducted on a series of 11 adult testis tissue samples, containing tubules with normal complete spermatogenesis and tubules with the pre-invasive carcinoma *in situ* (CIS), and two samples of prostate tissues were used as positive controls. The samples were obtained from the pathology archives of Rigshospitalet and had been snap-frozen, fixed in formalin or modified Stieve's fluid (GR-fixative) directly after orchidectomy, and subsequently embedded in paraffin.

### Ethics

All affected and unaffected family members gave their informed consent to participate in this study. The Regional Medical Ethics Committee approved first the genetic part of the study (Nr KF-01-26 5848) and subsequently the use of archived tissue samples for gene expression investigations (Nr H-2-2009-137).

### CNV detection by array CGH

DNA was purified from peripheral blood using the QuickGene-810 Nucleic Acid Isolation System (Fujifilm, Life Science-Products, Tokyo, Japan), by means of the

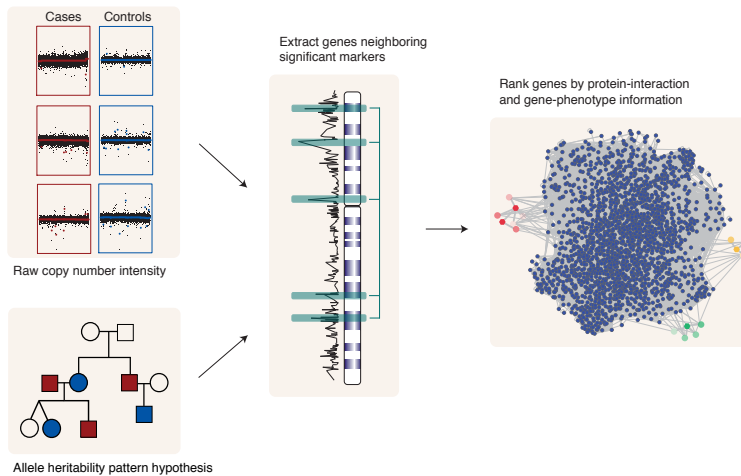
QIAmp 96 DNA, Blood kit (Qiagen, Inc., Chatsworth, CA), the Nucleo Spin, 96 Blood kit (Macherey-Nagel, Select Science Ltd., Corston, UK), or the DNA isolation kit for mammalian blood from Roche (Basel, Switzerland). Integrity of DNA was tested by gel electrophoresis and found to be high.

Array-CGH was performed with the human genome CGH microarray kit 244K and/or 185K (Agilent Technologies, Palo Alto, CA, USA). Sample DNA and normal male reference DNA (a pooled sample of DNA from the blood of 10 young healthy male military conscripts with a sperm count above 100 million / ml were used for hybridizations on the 244K chips according to the manufacturer's protocol, and as previously described,<sup>12</sup> whereas normal human male genomic DNA purchased from Promega was used as a reference pool on the 185K chips. Following hybridization, slides were washed and assessed for fluorescent signals using an Agilent G2565BA microarray scanner (Agilent Technologies). Image analysis was performed with Agilent Feature Extraction software with default settings, which includes spatial detrending and LOWESS normalization. To identify CNVs the marker-specific log-ratio output from the feature extraction software was subjected to segmentation by applying the circular binary segmentation (CBS) algorithm<sup>13</sup> with parameters set to 10000 permutations and a significance level alpha of 0.05. To avoid spuriously called change-points due to local trends in the data, break-points less than one standard deviation apart were removed. CNVs were required to span at least four markers. Additionally, to diminish the number of CNVs with low signal to noise ratio, CNVs with a mean absolute log<sub>2</sub>-ratio below 0.3 were discarded. Recurrent CNVs observed in several samples were defined as regions with a reciprocal overlap of at least 50% of the particular CNV sequences. The population frequency of deletions and amplifications at identified recurrent CNVs were estimated by manual inspection of frequencies listed in the browser of the 'Database of Genomic Variants' (DGV, version 9) which contains a collection of population studies of structural variants.<sup>14</sup>

### Prioritisation of candidate CNV loci by data integration

Due to the low number of samples, detected recurrent aberrations may not be disease causative. To alleviate the lack of information inherent in the sparse availability of samples and to pinpoint risk-associated genes within a set of candidate CNV loci, we employed a previously described disease gene prioritising method that utilizes sub-networks of protein-protein interactions and *a priori* gene-disease association information.<sup>11</sup> The method ranks each input gene by first performing a "virtual pull-down" of protein interaction partners from a database collection of experimentally established protein-protein interactions. The protein complex thus formed around each input gene is subsequently ranked based on how strongly the members of the protein complex are associated to the disease of interest, as well as the reliabilities of the protein-protein interactions. An important feature of the method is that the disease-gene association does not need to involve the exact disease of interest, since it is based on phenotype similarity.

To automatically select candidate genomic regions from CNV data we designed an additional first step to the method. Genes were selected and ranked if any nominally significant aCGH marker was found within a detected CNV locus ( $P < 0.05$ ),



**Figure 6.1** – Workflow of the systems biology method that ranks genes from CNV candidate loci. Samples are grouped into cases and controls according to a heritability hypothesis, and a test for association between copy number log-ratio signal and case-control status is performed. Second, genes within 5kb from a marker with indicated association are extracted. Third, a “virtual pull-down” of protein interaction partners is performed for each selected candidate gene. The protein complex thus formed is subsequently ranked based on the phenotypic similarity of the diseases associated to the members of the protein complex to the disease of interest, as well as the reliabilities of the protein-protein interactions.

and the marker was at maximum 5kb from a gene. The difference in copy number signal between cases and controls was assessed by performing a pooled two-sample Student t-test on the log-ratio signal of each marker. Marker positions were retrieved from the Agilent annotation file and gene positions from Ensembl using the NCBI v36 assembly. A variety of heritability patterns were tested within the pedigrees using different case and control configurations. This makes particular sense for the female samples as they may have a risk-associated CNV but cannot develop testicular cancer, and, furthermore, as it was not possible to make any inference of the inheritance pattern due to the low pedigree depth. The OMIM labels “Testicular tumors” (MIM:273300) and “Testicular germ cell tumor 1” (MIM:300228) were used as target phenotypes in the method. The workflow of the procedure is illustrated in Figure 6.1.

#### Validation of *RLN1* copy number using quantitative real-time PCR (qPCR)

The copy number of the *RLN1*-sequence at 9p24.1 was validated by qPCR analysis performed on the Mx3000P platform from Stratagene (Cedar Creek, Texas). The

protocol has been described previously.<sup>15</sup> Primers for the CNV at *RLN1* were designed using Primer3,<sup>16</sup> 2000). *GAPDH* was used as endogenous control gene for normalization, and expected product sizes of amplicons were 60 bp and 78 bp, respectively, demonstrated by gel electrophoresis. Primer sequences for *RLN1* were: forward 5'- CAT TTG GTG TGT AAG AAA ATA TTC TTT GT- 3' and reverse 5'- AGA AAG TGT CAT TTA CAG AAA CTA CAA TC -3' (DNA Technology A/S, Aarhus, Denmark), and primers for *GAPDH*: forward 5'- CTC CCC ACA CAC ATG CAC TTA -3' and reverse 5'- TTG CCA AGT TGC CTG TCC TT -3'. In brief, mixtures of forward and reverse primers were denatured for 3 min at 95°C and incubated on ice until use. DNA was isolated as described for aCGH above and reaction tubes contained 8-17 ng DNA, 15 µl Brilliant SYBR Green QPCR Master Mix (Stratagene), 7.0 µl primer mixture of *RLN1* (final conc.: Fw and Rev 400nM) or *GAPDH* (final conc.: Fw and Rev 100nM), and a total volume of 30 µl. Conditions for amplification were as follows: 1 cycle at 95°C for 10 min and 40 cycles at 95°C for 30 sec / 61°C for 1 min / 72°C for 1 min. The *RLN1:GAPDH* ratio was calibrated to a normal male reference control of DNA, as previously described for other genes.<sup>15,17</sup> All specimens were analysed in triplicate and the mean-ratio was used to infer integer copy number. The inference was done using SPSS (version 17), where histograms of the ratio values were used to derive intervals that correspond to certain copy numbers. For the determination of two copies the 2.5-97.5 percentiles were used, but due to the low number of cases with a CNV (one or three copies) the intervals for these were defined as the highest and lowest ratio values observed. Statistical analysis of the resulting allele frequencies was performed using R (version 2.9).

### Immunohistochemistry

Four anti-human relaxin polyclonal antibodies (Pab) for the *RLN1* and *RLN2* gene products were used: two rabbit Pabs for, Relaxin-H1 (FL-185, from Santa Cruz Biotechnology and RLX H1, kindly provided by G. Bryant-Greenwood), a rabbit Pab for Relaxin-H2 (RLX H2 provided by G. Bryant-Greenwood), a goat Pab targeting both Relaxin-H1 and -H2 (N-18, Santa Cruz Biotechnology), and a rabbit Pab raised against the relaxin receptor RXFP1 (LGR7 from Phoenix Pharmaceuticals Inc., Burlingame, CA, USA). Antisera to RLX H1 and RLX H2 provided by Dr. G. Bryant-Greenwood were purified using the Melon Gel IgG Spin Purification Kit (Pierce Thermo Fisher Scientific, Rockford, IL, USA) according to the manufacturer's protocol. The purified antibodies were aliquoted and stored at -20°C until further use.

Immunohistochemical staining was performed on deparaffinised and rehydrated sections of adult human testis (with and without pre-invasive testicular carcinoma in situ) according to a standard indirect peroxidase method, using epitope retrieval in a microwave oven. Prostate tissue samples were used as positive controls, since all studied proteins have been previously described in this tissue.<sup>18-20</sup> Incubation with the primary antibodies was carried out O/N at 4°C. As negative control a serial section from each block was incubated with a dilution buffer (TBS) alone or with normal pre-immune rabbit serum (DakoCytomation, Glostrup, Denmark). As a secondary layer, biotinylated goat anti-rabbit or donkey anti-goat anti-

bodies were used, followed by peroxidase-conjugated streptavidin complex (Invitrogen, Carlsbad, CA, USA). The bound antibody was visualized using aminoethyl carbazole (AEC kit from Invitrogen).

## Results

### Recurrent CNV analysis

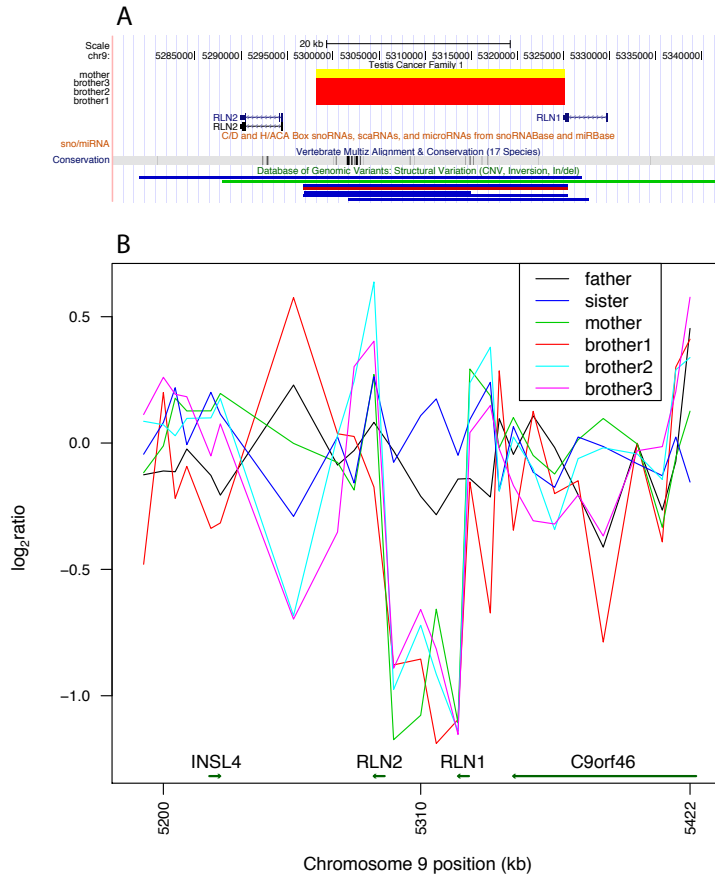
We performed genome-wide CNV analysis of constitutive DNA (blood cells) from 16 members of four testicular cancer-prone families using an Agilent 244k and/or 185k aCGH platform (Supplementary Table 1). Members of family 1 (three brothers with GCT) had previously been analysed by low-resolution cytogenetic analysis, but no abnormalities were then found.<sup>21</sup> We first analysed each individual family to find recurrent CNVs that affected all of the disease-afflicted family members. The number of such recurrent CNVs per family was eight, eight, five and zero respectively (Table 6.1 and Supplementary Table 2). After inspection of the 'Database of Genomic Variants', three of the 20 family-recurrent CNVs turned out to be rare, defined as a minor allele frequency (MAF) below 3%: two CNVs within family 1 located at 9p24.1 and 16q21, and one CNV within family 2 at 1q21.3 (Table 6.1). The CNV at 9p24.1 comprised a 26kb-long heterozygous loss, detected in all three GCT afflicted brothers of family 1 as well as the mother and was demonstrated by four aCGH markers with the most downstream marker located within the exon of *RLN1*. The CNV had identical breakpoints across all four affected samples and the log<sub>2</sub>-ratios were approximately - 1 for all four markers in all four samples, indicating a heterozygous loss (Figure 6.2). 16q21 contained a 237kb long region with a gain, approximately 500kb downstream of the closest gene *CDH11*, where – apart from the three brothers in family 1 – the father and sister also harboured the CNV. The third CNV was found in family 2 where both GCT afflicted brothers had a 22kb long loss at 1q21.3, covering exons of *SEMA6C*.

Furthermore, a test for CNVs shared between families revealed eight losses and five gains that occurred in at least five of the 16 individuals and were found within more than one family (Supplementary Table 2). All of the eight losses were common CNVs with high frequencies in population studies. However, one of the gains, a 17.1kb long CNV at 9q34.3 covering exons of *NOTCH1*, had a somewhat lower estimated MAF of about 3%. The individuals carrying the CNV were one of the affected brothers, the sister and the mother of family 1 as well as two brothers with TGCT in family 4.

### Prioritisation of candidate CNV loci by integrative systems biology

We employed a data integrative gene ranking method (see Methods) on the genome-wide CNV data from each individual family using different heritability hypotheses and two different testis cancer OMIM identifiers (MIM:300228 and MIM:273300). This analysis identified one protein-complex with a posterior probability of 0.95, which was considerably higher than that of any other complex (Table 6.2). In this context the posterior denotes the probability that a protein complex is associated to TGCT given the presented data, as compared to the probability of





**Figure 6.2 – (A)** CNV at 9p24.1 covering one exon of *RLN1* and extending towards *RLN2*. A loss was observed in all three GCT afflicted brothers as well as the mother in family 1. **(B)** Marker-specific signal for the loss at 9p24.1. Y-axis indicates  $\log_2$ -ratio of the marker-intensity signal between sample and a pool of references, and the X-axis denotes genomic position.

Family	CNV	Type	Gene relation	Length (kb)	DGV <sup>a</sup>	Aff. cases <sup>b</sup>	Aff. controls <sup>c</sup>
1	chr9:5298149..5325065	Del	RLN1, exon	26.9	1-1.5%	3/3	1/3 (mother)
1	chr16:62882631..63119309	Amp	CDH11, 500kb downstream	236.7	0%	3/3	1/3 (father)
2	chr1:149369523..149391633	Del	SEMA6C, exons	22	0%	2/2	0/2
All 4 fam. <sup>d</sup>	chr9:138516970..138534059	Amp	NOTCH1, exons	17.1	2-3%	3/10	2/6 (mother, sister)

**Table 6.1** – CNVs recurrent among the GCT afflicted family members and with minor allele frequency (MAF) less than 3%. Supplementary Table 2 includes CNVs with higher MAF and specifies the individuals harbouring the CNVs. <sup>a</sup>Database of Genomic Variants, <sup>b</sup>CNV afflicted cases / all investigated cases of specified family, <sup>c</sup>CNV afflicted controls / all investigated controls of specified family, <sup>d</sup>CNVs found in at least five individuals from more than one family.

any other evaluated complex. The posteriors were calculated for each specific family and heritability hypotheses, and were not corrected for this, resulting in that the sum of the posterior probabilities for all evaluated complexes exceeds one. The input gene that resulted in the highranked complex was *RLN1* from candidate CNV locus 9p24.1, and the ranked complex consisted of the peptide hormones *RLN1* and *RLN2*, and the receptors *RXFP1* and *RXFP2*. This complex ranked high, as *RXFP2* is a receptor of *INSL3*, that has previously been associated with phenotype terms such as cryptorchidism and infertility, which are risk factors for testicular cancer, and whose text descriptions are similar to that of testicular tumours. The finding of *RLN1* by the disease gene ranking method was in concordance with the analysis of recurrent CNVs described above, and highlighted 9p24.1 among the CNV loci as being the most likely to be associated with testicular cancer. Among the hundreds of ranked genes none of the other genes identified by the recurrent analysis achieved a high posterior probability. This could be due to incomplete interactome information or due to a lack of previous gene-phenotype relations suggestive of an association with germ cell development or function.

Family	CNV heritability hypothesis <sup>a, b</sup>	HUGO	Locus	Posterior probability <sup>c</sup>	MIM <sup>d</sup>
1	br1, br2, br3, mom	RLN1	9p24.1	0.95	300228
2	br1, br2, mom	MALT1	18q21.32	0.40	273300, 300228
4	br1, br2, mom, uncle	HUWE1	Xp11.22	0.18	273300
1 (185k)	br1, br2, br3, mom	HUWE1	Xp11.22	0.15	273300
4	br1, br2, uncle	HUWE1	Xp11.22	0.14	273300
4	br1, br2	TRGJP, TRGJ1, TRGC1	7p14.1	0.12	273300, 300228
4	br1, br2	CCDC154	16p13.3	0.12	273300, 300228
4	br1, br2	ILIRAPL1	Xp21.3	0.12	273300, 300228

**Table 6.2** – Top-ranked genes by the disease-gene systems biology algorithm. *RLN1* shows a clear phenotype association signal as indicated by the posterior probability. <sup>a</sup>CNV heritability hypothesis indicates which family members were treated as cases harbouring a minor CNV allele, <sup>b</sup>br – brother, <sup>c</sup>Genes with a posterior probability above 0.1 are presented. The sum of probabilities exceeds one, as they were calculated specifically for each family and heritability hypothesis, <sup>d</sup>Identifier in the ‘Online Mendelian Inheritance in Man’ database.

### In silico analysis of *RLN1* genomic region

A search for functional elements in the non-coding region of the CNV, downstream of *RLN1*, did not identify any promoters or miRNAs, but a computationally predicted 197bp pseudogene listed in Ensembl. As recently described, pseudogenes may act as competitors for miRNA,<sup>22</sup> but the possible function of this potential pseudogene was not further explored in this study.

### qPCR validation of *RLN1*-copy number in patients with familial GCT

qPCR validation of the *RLN1*-copy numbers in family 1 showed 100% agreement with the aCGH data, the three sons and the mother demonstrating one copy, and the father and sister two copies of the gene. We further tested eight patients with familial TGCT from five other families (with two affected males in each family), as well as three healthy family members, and all individuals showed a normal genotype of two *RLN1* copies.

### *RLN1* validation by a case-control association study

The potential disease association of the CNV at *RLN1* was further investigated by performing qPCR on an external validation set of a total of 306 individuals, 106 patients with sporadic TGCT and 200 young male conscripts after removal of six samples that were outside the reference intervals. Two patients demonstrated one copy of *RLN1* and one had three copies. In the control group three individuals had one copy and two had three copies (Table 6.3). None of these 5 control individuals had a history of cryptorchidism or any sign of gonadal dysfunction and all had normal hormone profiles. Four of these men had normal sperm concentration ranging from 53-158 mill/ml, but one demonstrated only 18 mill/ml. A Fisher's exact test on the aberration frequencies of *RLN1* indicated no significant difference between the cases and controls ( $P = 1$ ). The odds-ratio for a deletion was 1.26, suggesting a weak risk of disease association, however the sample size was too small to make any reliable conclusion. The distribution of *RLN1*-copy number frequencies along with odds-ratios and p-values is presented in Table 6.3.

Copy number	Controls (n=200)		TGCT patients (n=106)	
	N	%	N <sup>a</sup>	%
2	195	97.5	103	97.2
1	3	1.5	2	1.9
3	2	1	1	0.9
Total MAF <sup>b</sup>	5	2.5	3	2.8

**Table 6.3** – *RLN1* copy number state frequencies in the qPCR association study.

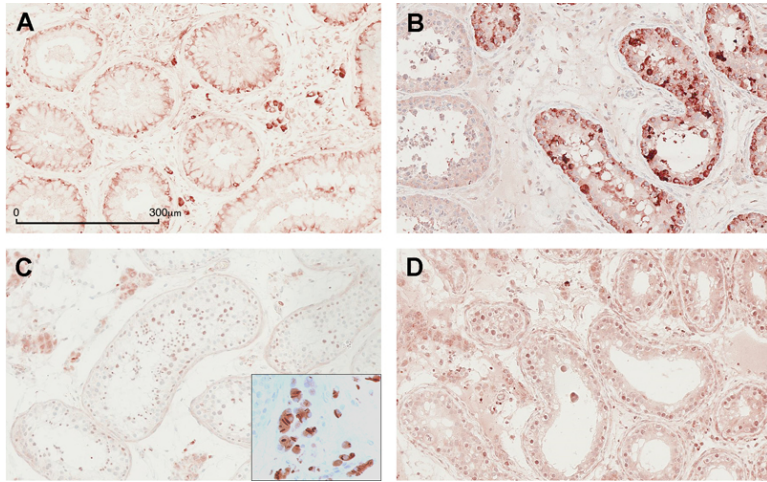
<sup>a</sup>The three brothers of family 1 were included as a single case. <sup>b</sup>Minor Allele Frequency.

### Expression of RNL1, RNL2 and their receptor RXFP1 in the human testis and carcinoma *in situ*

To investigate the biological relevance of the discovered *RLN1* copy number variation, we performed a pilot immunohistochemical study of the expression of the *RLN1* and *RLN2* peptide products, Relaxin-H1 and -H2, and their receptor in 11 different specimens of the human testis containing tubules with normal complete spermatogenesis and tubules with the pre-invasive carcinoma *in situ* (CIS). The results varied slightly depending on the tissue fixative, but all antibodies showed a good positive reaction in the glandular epithelium of the prostate that was used as a positive control (Supplementary Figure 1), in agreement with previous publications.<sup>18–20</sup> After further validation of the specificity of the antibodies by western blotting and competition assay with N-18 antibody (not shown here), we performed repeated staining in several fixatives and frozen tissues. Only the antibody that could not distinguish the two relaxins (N-18) and the antibody against the receptor (RXFP1) showed bands of expected size on the western blot, thus we show in Figure 6.3 only examples of stainings obtained with these two antibodies. Despite a poor performance in western blotting, the two antibodies to Relaxin-H2 gave quite consistent results in Leydig cells in different fixatives. The observations are therefore very preliminary, but from the overall staining pattern we deduce that both relaxins were most likely present in the interstitial compartment, as shown by strongly positive staining by all three antibodies (Figure 6.3). In the intratubular compartment it was more difficult to interpret the pattern, and it remains to be confirmed by additional studies, but we observed either Relaxin-H1 or Relaxin-H2 in CIS cells in 10/11 specimens (Figure 6.3 A, B, C). Relaxin-H2 was predominantly present in Sertoli cells but the latter staining was only visible in 2/11 specimens and is thus uncertain (Figure 6.3B). The receptor RXFP1 was most abundant in the cytoplasm of Leydig cells, strongly marking Reinke crystals, but was also detected in a subset of spermatids for all 11 stained samples (Figure 6.3C).

### Discussion

In a search for gene candidates predisposing to testicular germ cell tumours (TGCT), we applied an integrative systems biology algorithm to a genome-wide CNV data set on 16 individuals from four TGCTprone families. Familial TGCT is relatively rare in Denmark, and we could therefore only include a limited number of families with Danish ethnicity. Small-scale family-based CNV association-studies are seldom powerful enough to identify disease susceptibility genes. Recurrent CNVs may be family-specific harmless genomic alterations or may be common polymorphisms with no disease-associated risk. CNVs may also span a large genomic region, making it difficult to determine the critical locus within a CNV. Additionally, if a CNV does not have high penetrance there may be family members carrying potentially harmful CNVs, but without any observed phenotypic conditions. To restrict the number of possible candidates, we extended a recently published candidate disease-gene ranking method<sup>11</sup> to automatically extract candidate regions from CNV data. A somewhat similar bioinformatics platform for candidate gene prioritization that also makes use of network information, Prioritizer,



**Figure 6.3** – Examples of immunohistochemical localization of Relaxin-H1/2, and the receptor RXFP1 in the human testis and pre-invasive testicular carcinoma *in situ* (CIS). (A) Relaxin-H1/2: positive reaction in Leydig cells and Sertoli cells, (B) Relaxin H1/2 shows a strong, predominantly cytoplasmic reaction in CIS cells, but only weak reaction in Sertoli cells and adjacent normal tubules, (C) RXFP1 staining in Leydig cells, a subset of spermatocytes and spermatids. An insert shows a higher magnification of RXFP1-positive Leydig cells with strongly stained Reinke crystals, (D) RXFP1- positive CIS cells with weak, predominantly nuclear staining.

has recently been applied in the context of CNV susceptibility loci,<sup>23</sup> but it does not incorporate *a priori* disease-gene annotations or utilize phenotypic proximities that may be critical to improve detection in a small-scale CNV study.

The disease-gene ranking method in combination with the statistical analysis of loci from the aCGH identified one CNV at 9p24.1 within family 1 with a strong significance. The recurrent CNV analysis identified three additional CNVs that were present in disease-afflicted family members and were relatively rare polymorphisms (at *CDH11*, *SEMA6C* and *NOTCH1*), but the integrated analysis incorporating interactome data and previous gene-phenotype relations, reported a low potential for these loci to be associated with testicular cancer.

Copy number variations at *RLN1* have previously been demonstrated in four individuals from HapMap<sup>24</sup> revealing a heterozygous deletion at this locus: three persons of European ancestry (CEU) comprised of two mothers (NA12763, NA12717) and the son of one of them (NA12707), as well as a male of Asian ancestry (CHB, NA18608).<sup>25–27</sup> Furthermore, 21 individuals from a collection of 2026 samples were annotated to have a deletion in this region,<sup>28</sup> indicative of the deletion being relatively rare, with a minor allele frequency of about 1%. It is probable that the deletion is caused by the segmental duplications (SDs) found to overlap the CNV, as SDs are known to define hotspots of chromosomal rearrangements.<sup>29,30</sup> It is also

evident that this part of the genome has been subjected to duplication in its evolutionary history shown by the presence of four insulin-like genes located next to each other, *INSL6*, *INSL4*, *RLN2* and *RLN1*, all highly homologous.<sup>31,32</sup>

Validation by means of qPCR confirmed the presence of a heterozygous deletion at *RLN1* in the three brothers and the mother in the first GCT family. External validation, however, did not show a significantly higher frequency of deletions at *RLN1* in patients with sporadic TGCTs compared to healthy controls. This finding suggests that the aberration alone may infer a relatively low risk of TGCTs, but could possibly act in concert with other genetic or environmental factors. Thus the possible involvement of this CNV in the pathogenesis of family 1 members cannot be excluded. This would be in concordance with the fact that large linkage studies have failed to identify a single locus explaining a large fraction of the total risk of GCT, suggesting multiple susceptibility loci.<sup>3</sup> The fact that relatively few individuals with TGCT have a deletion at *RLN1*, could indicate that there is a plethora of rare susceptibility variants and that certain combinations of these may all result in a high penetrance.

The question could be raised whether the control group of the validation cohort was optimal since it was constituted of young males aged 18-21 years, who may develop testicular cancer later in life, but the lifetime risk of testicular cancer in this age group is around 0.6%.<sup>33</sup> Additionally, most of the subjects had good reproductive parameters (concentration of spermatozoa in semen samples and normal levels of inhibin B and testosterone), which may even further decrease the risk, since this cancer is associated with poor testis function.<sup>34,35</sup> We are therefore confident that this group reasonably well reflects the background population. The five control individuals who displayed CNV in the *RLN1* locus were healthy and had normal testis function, suggesting the possible redundancy of this gene. On the other hand, we have not in any subject (patient or control) identified a homozygous deletion of the *RLN1* locus, which may suggest lethality of such a genotype, but to confirm that a much larger cohort would need to be investigated.

As opposed to Relaxin-H1, Relaxin-H2 is a well-known hormone in the context of female reproduction, where it is commonly known as one of the pregnancy hormones. It is produced in large quantities by ovarian corpus luteum, but it also acts locally in the decidua, placenta, endometrium and the mammary gland (2009). Relaxins have not been as extensively studied in human male<sup>36</sup> and both Relaxin-H1 and H2 have only been described in the prostate<sup>19,37</sup> and in seminal fluid, where it is thought to affect the motility of spermatozoa.<sup>38</sup> The available knowledge on Relaxin-H1 is very scarce, and some consider it a redundant effect of a recent genomic duplication, as it is only present in primates.<sup>31</sup> We made preliminary expression examinations of a putative protein product of *RLN1* (Relaxin-H1), the closely related peptide Relaxin-H2, encoded by *RLN2*, as well as its receptor *RXFP1*. Our findings in this study, although very preliminary, suggest that relaxins may also play a role in the human testis, as we detected their presence both in the interstitial compartment and the seminiferous epithelium. However, we have to stress that the expression of relaxins in the testes appears in general much lower than that in the prostate and that the available antibodies may cross-react with other proteins, as suggested by western blotting experiments. The results indicate that

relaxin peptides are predominantly produced in Leydig cells and to a lesser extent in Sertoli cells, which is consistent with reports from animal studies,<sup>39,40</sup> however rodents produce only one relaxin hormone, considered as the equivalent of human Relaxin H2. The relaxin signalling system has not previously been associated to, or studied, in testicular cancer, but only reported to be involved in prostate cancer.<sup>20</sup> We observed quite consistently the presence of at least one, perhaps both, of Relaxin-H1 and -H2, and their receptor in CIS cells. This is supported by the presence of *RLN1* transcripts in micro-dissected CIS cells as detected in our previous study of global gene expression profile of CIS.<sup>41</sup> However, at this point it is unclear, because the most reliable immunohistochemical data were obtained using an antibody that cannot distinguish the two peptides. Furthermore, the presence of relaxins in male germ cells has never previously been reported. We detected the relaxin receptor *RXFP1* in Leydig cells and in spermatids, the latter corroborated by studies in rats.<sup>42</sup> Supportive of these indications are results from relaxin null male mice displaying decreased sperm maturation and increased cell apoptosis in the testis.<sup>43</sup> Taken together the evidence is not yet conclusive and awaits development of new antibodies with better specificity, as also suggested by others.<sup>36</sup> Only then the preliminary observations of this pilot investigation can be confirmed in a larger study.

In summary, we here presented and applied a method that extracts candidate disease-causing genes from copy number variation association studies. The incorporation of interactome and phenome data in conjunction with statistical tests to assess different heritability patterns makes it especially useful for small studies, such as family studies of moderately rare diseases, where the small number of samples can limit the possibility to make significant findings by solely using the CNV data. In this study, employment of this systems biology method identified a heterozygous deletion affecting the *RLN1* gene. This CNV is apparently not associated with an increased risk of TGCT, but we cannot exclude that it may contribute to a risk in some rare cases of familial cancer, probably in conjunction with other genetic or environmental factors. The results also indicate a possible redundancy of the *RLN* loci in humans but nevertheless warrant further studies of the expression patterns of *RLN1*, *RLN2* and the *RXFP1* receptor in adult testis as well as during the development of the reproductive organs, in order to elucidate the biological role of these peptides in the human male.

### Acknowledgements

We thank Dr. Gillian Bryant-Greenwood for the provision of antibodies, Drs Gedske Daugaard and Niels Jørgensen for their assistance with the inclusion of patients and controls, Marlene Dalgaard for providing diagnostic information, John E. Nielsen for help with immunohistochemistry, Dr Anne Jørgensen for help with western blotting and Inger D. Garn and Hanne O. Mogensen for technical assistance. The authors are grateful to Professor Per Guldberg for providing access to the microarray processing facility at the Danish Cancer Society. The work was supported by grants from the Villum Kann Rasmussen foundation, Aase and Ejnar Danielsen Foundation, and the Danish Cancer Society.

## 6.3 Supplementary Material

Available in Appendix A.





---

## Chapter 7

# Genome-wide association study of men with testicular dysgenesis syndrome

---

### 7.1 Abstract

The testicular dysgenesis syndrome (TDS) is a hypothetical common entity that links testicular germ cell cancer, cryptorchidism, and some cases of hypospadias and male infertility with an impaired development of the testis. Incidence of these disorders has increased over the last decades and testicular cancer is now affecting 1% of Danish and Norwegian males. In order to identify genetic variants that span across these phenotypes, we conducted a genome-wide association study and combined it with a systems biology approach, which increases the significance of the markers identified. Significant association was found at *KITLG* in discovery and replication, but only to testicular cancer. Markers at *TGFBR3* and *BMP7* found with system biology tools showed association across phenotypes both in discovery and replication, indicating a role of the *TGF $\beta$*  superfamily signaling pathway in the pathogenesis of TDS.

## 7.2 Manuscript

### A genome-wide association study of men with testicular dysgenesis syndrome

Marlene D. Dalgaard<sup>2,†</sup>, Nils Weinhold<sup>1,†</sup>, Daniel Edsgård<sup>1,†</sup>, Jeremy D. Silver<sup>4</sup>, Tune H. Pers<sup>1,3</sup>, Niels Jørgensen<sup>2</sup>, Anders Juul<sup>2</sup>, Thomas A. Gerds<sup>4</sup>, Alexander Giwercman<sup>5</sup>, Yvonne Lundberg Giwercman<sup>5</sup>, Gabriella Cohn Cedermark<sup>6</sup>, Helena E. Virtanen<sup>7</sup>, Jorma Toppari<sup>7,8</sup>, Gedske Daugaard<sup>8</sup>, Thomas Skøt Jensen<sup>1</sup>, Søren Brunak<sup>1</sup>, Ewa Rajpert-De Meyts<sup>2</sup>, Niels E. Skakkebæk<sup>2</sup>, Henrik Leffers<sup>2</sup>, Ramneek Gupta<sup>1,#</sup>

<sup>1</sup> Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

<sup>2</sup> Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.

<sup>3</sup> Institute of Preventive Medicine, Copenhagen University Hospital, Centre for Health and Society, Copenhagen, Denmark.

<sup>4</sup> Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark

<sup>5</sup> Department of Clinical Sciences, Molecular Genetic Reproductive Medicine, University, Malmö, Sweden.

<sup>6</sup> Department of Oncology-Pathology, Karolinska University Hospital, Stockholm, Sweden.

<sup>7</sup> Departments of Physiology and Pediatrics, University of Turku, Turku, Finland.

<sup>8</sup> Department of Oncology, Rigshospitalet, Copenhagen, Denmark.

† These authors contributed equally

# Correspondence should be addressed to ramneek@cbs.dtu.dk

### Introduction

Infertility, cryptorchidism, small testis size with poor semen quality and hypospadias are relatively common among men in Western countries and are all risk factors for testicular germ cell cancer (TGCC)<sup>1,2</sup> which has been increasing in incidence and is the most common malignancy of young men in many countries.<sup>3</sup> Caucasian men are at particularly high risk of testicular cancer, currently affecting approximately 1% of the Danish and Norwegian male populations.<sup>4</sup> It has been proposed that most cases of TGCC and cryptorchidism, and some cases of hypospadias and decreased spermatogenesis (excluding those unrelated to development of the fetal testis), may be symptoms of a testicular dysgenesis syndrome (TDS) that originates from perturbations during fetal development.<sup>2,5</sup>

Recent progress in basic and epidemiological research on testis cancer has provided evidence that TGCC, in spite of its occurrence in young adult men, has its

origin in fetal germ cells with stem cell characteristics,<sup>6,7</sup> the so-called intratubular carcinoma *in situ* germ cells (CIS).<sup>8</sup> According to a widely tested hypothesis, these cells are derived from gonocytes that failed to mature to spermatogonia due to dysfunction of Sertoli, peritubular, and Leydig cells, whose function is essential for differentiation of fetal germ cells.<sup>6,9</sup> However, the invasive potential of CIS cells is not attained until after the pituitary-gonadal axis is activated during puberty.

Little is known about the causes of TDS, although environmental factors appear to play an important role, as suggested by the rapidly increasing incidence of TGCC that has occurred over a couple of generations.<sup>3</sup> The prevalence of mild forms of cryptorchidism and poor semen quality also appear to be increasing.<sup>10,11</sup> There is evidence of a genetic contribution, as brothers and fathers of patients with TGCC have a significantly increased risk of TGCC.<sup>12</sup> Marked ethnic differences in risk of TGCC amongst men living in the same areas also indicate a noticeable genetic component in the aetiology of the disease.<sup>3</sup> Recently, three genome-wide association studies (GWAS) identified several genetic loci that predispose to TGCC (*KITLG*, *SPRY4*, *TERT-CLPTM1L*, *ATF7IP* and *DMRT1*). Altogether, these loci can potentially explain up to 13% of the heritability.<sup>13-15</sup> The variants with highest effect size were found at 12q21, implicating *KITLG*/*KIT* signaling as a pathway involved in TGCC susceptibility.

Based on our hypothesis that individuals with TDS may be genetically predisposed to yet unknown environmental factors affecting interrelated pathways during testis development, we conducted a novel GWAS where we grouped cases with at least one of four TDS phenotypes: TGCC, cryptorchidism, hypospadias, or infertility with low sperm concentration. The GWAS was conducted on a discovery cohort of 926 Danish men, comprised of 439 healthy controls, and 488 patients affected by at least one TDS phenotype. The cohort was carefully selected with the help of detailed patient records that provided background information on individuals, and additional phenotype characterization including semen analysis. Candidate markers from the discovery stage were subsequently assessed in an independent replication cohort of 671 Nordic men. The markers were selected based on three approaches: (1) Conventional single-marker association, (2) an integrative systems biology approach based on augmenting the GWAS data with several complementary types of data, and (3) protein-complex based pathway analysis, which examined if groups of related genes in the same functional pathway were jointly associated with the trait of interest.

Systems biology approaches that exploit the complementarity of heterogeneous types of data have recently been successfully applied in GWA studies to detect associations that may be missed by single marker association.<sup>16-18</sup> Here we used two such approaches. First, we adapted the Metaranker tool,<sup>19</sup> to integrate gene-phenotype associations from several complementary data sources, i.e., targeted knockout experiments in mice, transcriptome time-series expression studies of the developing testis, and protein-protein interactions.

In a second approach we tested protein complexes for association to TDS, therefore using a pathway-based method to examine multiple markers from genes interrelated by prior biological knowledge, a type of methodology that recently has gained attention.<sup>20</sup>

With the above approaches we were able to: (1) identify pleiotropic risk variants for TDS, i.e., variants associated to all of the phenotypic subtypes, (2) confirm the risk factors in the *KITLG* gene found by the two recent genome-wide association studies on TGCC, and, (3) identify genes or protein complexes through integrative systems biology approaches that would not be found by single marker association.

## Results

In this study we genotyped a discovery cohort comprised of 439 controls and 488 cases affected by at least one of four TDS sub-phenotypes: testicular germ cell cancer, cryptorchidism, hypospadias or low semen concentration. Samples were analyzed using the Affymetrix Genome-Wide Human SNP Array 6.0. After application of stringent quality control criteria, we compared the genotype frequency distributions between cases and controls for 600,798 markers using the MAX test,<sup>21</sup> which allows simultaneous investigation of three plausible inheritance models: additive, dominant and recessive inheritance. We observed only marginal confounding due to population stratification or other biases as indicated by the genomic inflation factor ( $\lambda$ ) of 1.019 and the quantile-quantile (qq) plot of observed versus expected  $\chi^2$  test statistics (Supplementary Figure 2).

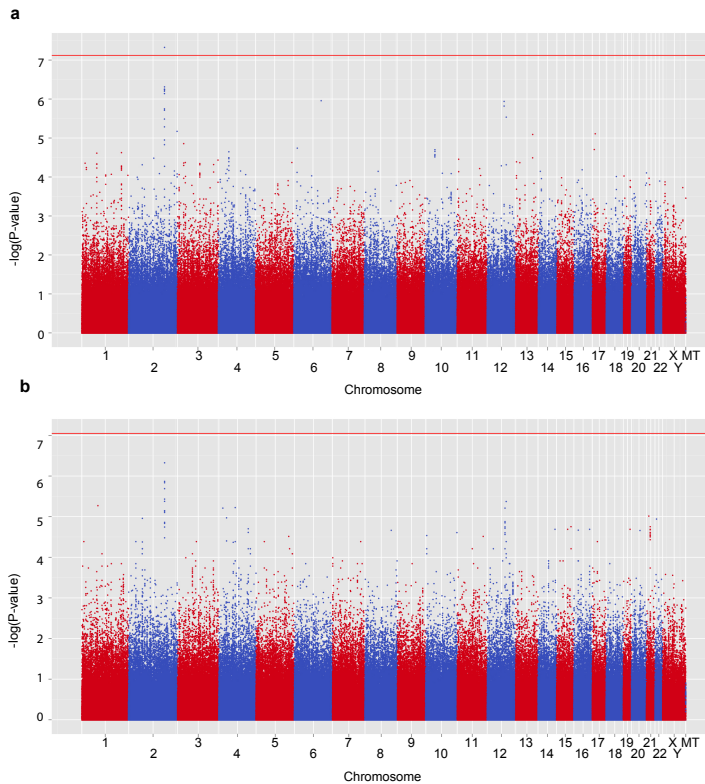
### Single marker association

We identified a strong association with all phenotypes of TDS ( $P < 10^{-6}$ ) for nine markers, within the same linkage disequilibrium (LD) block, at 2q31.1, with the strongest association for rs17198432 (Figure 7.1). These markers were located between two genes, *ATP5G3*, an ATP synthase, and *lunapark (LNP)*, a gene recently found to be co-regulated with genes from the nearby *HOXD* gene cluster.<sup>22</sup>

The subset of 212 TGCC cases in our cohort was investigated in a separate single marker analysis in order to compare the corresponding TGCC specific results to those of three recent GWA studies on TGCC.<sup>13-15</sup> The markers at *HOXD* remained the most significant also for this subset of the case cohort, but were closely followed by markers at *KITLG*. The most significant marker associated to *KITLG* was ranked 12th among all markers ( $P = 1.17 \times 10^{-5}$ ). Furthermore, in the TGCC group the observed genotype frequencies of the *KITLG* markers were comparable to those previously reported, (Table 7.1).

### Associations by meta-evidence based integrative systems biology

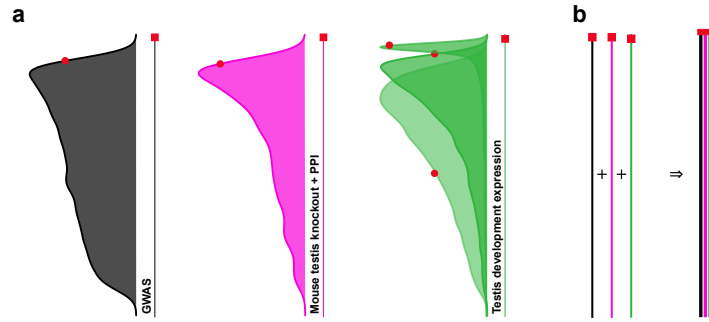
Despite the successful endeavor of conventional GWAS identifying single markers associated with traits, it is likely that part of the heritability is hidden in variants with lower effect sizes, which a given single study may not have the power to detect. To address this issue we employed two complementary approaches to select candidate markers to be included in the replication study: integration of gene-phenotype based data and a pathway-based analysis. These methods utilize the biological knowledge of the disease and of gene relations that are available in public databases, to improve the power.



**Figure 7.1** – Manhattan plots showing the association between all SNPs and (a) TDS, and, (b) the subset of cases with TGCC.

### Integration of gene-phenotype associations

For each data type all genes were ranked based on their strength of association to the TDS phenotype. These data type-specific ranked lists of genes were then combined into a final ranked list of genes. In order to avoid high ranking of genes that were only supported by a single data type, we used a modification of the Metaranker<sup>19</sup> tool by using the sum rather than the product of the individual gene-ranks. Thus, all human genes were ranked based on a combination of microarray gene expression time-series studies of the developing fetal testis in mouse and human,<sup>23–25</sup> protein-protein interaction data,<sup>26</sup> targeted mice knock-outs resulting in developmental defects of the testis (MGI), and gene-scores based on the single marker analysis of this GWAS (Figure 7.2; see Materials and Methods for details). From this analysis we selected markers from 14 top-ranked genes, many of which are active during early development (Supplementary Table 3 and 4).



**Figure 7.2** – Integration of GWAS data with heterogeneous data types. The rationale behind the method is to prioritize genes if several data types show mildly pronounced associations to the phenotype, thereby identifying genes that would not have been found with GWAS analysis alone. *TGFBR3*, which was selected by this method and whose SNP markers were further validated, is shown in red. The position of *TGFBR3* is marked by red circles in the distributions of p-values, and by red squares in the ranking of each data type layer. Three data types were used; (i) single marker GWAS results of this study (gray), (ii) targeted mutations from the Mouse Genome Informatics (MGI) database, filtered for TDS phenotypes, and ranked by their enrichment in protein-protein complexes (pink), (iii) differential expression in the fetal testis of mouse and human (green). (a) A distribution of the p-values of all human genes is shown for each of the data types. The p-value ranges from 0 to 1, with 0 at the top. The p-value associated to each gene is converted to a rank, such that each human gene is assigned a rank. The vertical line to the right of each distribution represents the ranking of all genes, where the top corresponds to rank 1. (b) The gene-ranks of each individual data type are combined to a final meta-rank of each gene. As an example, *TGFBR3* had the ranks 238, 42 and 408 among all human genes in the three data types: GWA, targeted mutations with protein-protein interaction enrichment, and differential expression in the developing testis, respectively. After combination of these data type specific ranks, *TGFBR3* was ranked 3rd among all genes.

Gene	Marker <sup>a</sup>	Selection method <sup>b</sup>	Phase	RA	Control RAF <sup>c</sup>	Per allele	Odds Ratio		$P_{unadj.}$	$P_{adj.}$
							Optimal genetic model <sup>d</sup>			
<i>KITLG</i>	rs1352947	TGCC	Discovery	T	0.81	1.56 (1.19-2.05)	1.52 (1.18-1.97)	add	$3.1 \times 10^{-3}$	1
	C/T		Replication		0.82	2.11 (1.48-3.03)	1.93 (1.39-2.69)		$8.5 \times 10^{-5}$	$2.0 \times 10^{-3}$
<i>TGFBR3</i>	rs12082710	ISB	Discovery	T	0.58	1.35 (1.10-1.65)	1.77 (1.33-2.36)	rec	$2.4 \times 10^{-4}$	1
	T/C		Replication		0.59	1.27 (0.99-1.63)	1.52 (1.08-2.15)		$1.6 \times 10^{-2}$	$3.8 \times 10^{-1}$
<i>BMP7</i>	rs388286	Pathway	Discovery	C	0.47	1.34 (1.10-1.63)	1.36 (1.11-1.67)	add	$2.3 \times 10^{-3}$	1
	C/T		Replication		0.47	1.29 (1.01-1.66)	1.28 (1.01-1.62)		$4.1 \times 10^{-2}$	$9.9 \times 10^{-1}$
<i>HOXDx</i>	rs17198432	TDS	Discovery	A	0.07	2.31 (1.66-3.26)	2.58 (1.82-3.70)	dom	$4.7 \times 10^{-8}$	$2.2 \times 10^{-2}$
	C/A		Replication		0.11	0.96 (1.43-0.64)	0.97 (0.64-1.48)		$8.7 \times 10^{-1}$	1

**Table 7.1** – Markers and genes with a possible association to the TDS phenotype. <sup>a</sup>The marker with the lowest p-value in the discovery cohort, among the markers tagging a gene, is presented. <sup>b</sup>Four different approaches were used for the selection of markers: TDS: single marker GWA on all TDS sub-phenotypes; TGCC: single marker GWA on the TGCC subset of cases; Pathway: aggregated effect in pathways; and, ISB: integrative systems biology by combing evidence of association from several data types. <sup>c</sup>Risk allele frequency among controls. <sup>d</sup>The genetic model with strongest association among the models tested by the MAX test (selected on discovery cohort), add: additive, rec: recessive, dom: dominant.

### Pathway analysis

Joint analysis of sets of markers may identify aggregated effects that are significant at a pathway level. Such analysis is motivated by the hypothesis that a set of genetic variations can perturb different components of a certain biologically functional entity, such as a signaling pathway or a protein complex, and they all affect the possibility of disruption of that biological entity. Thus, we derived protein complexes from protein-protein interaction data, and tested each complex for significant enrichment of TDS-associated SNPs. The most significant set of interacting proteins was identified as a complex consisting of members of the TGF $\beta$ superfamily: BMPER, BMP2, BMP7, BMP6 and BMP4 (Supplementary Table 6).

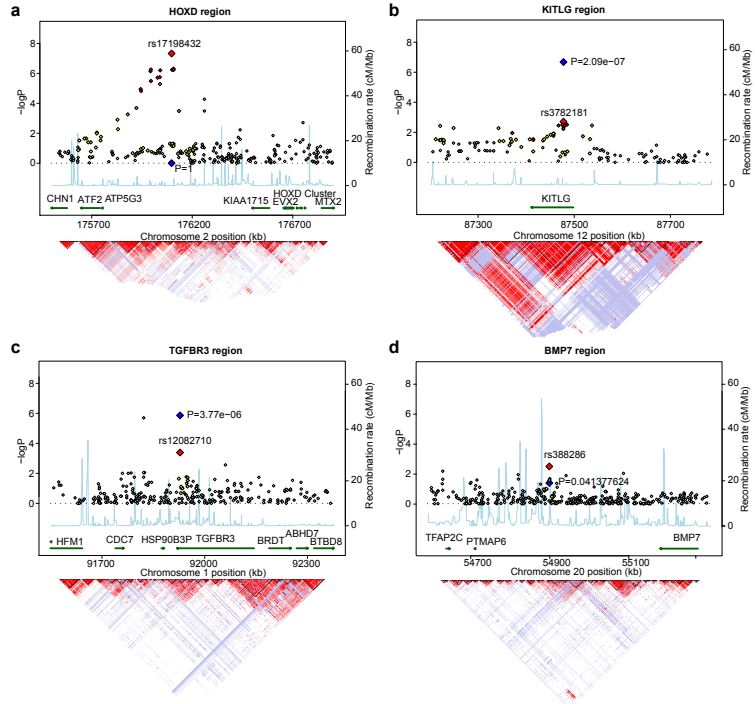
In summary, altogether 39 SNPs were put forward to a replication stage (Supplementary Table 2), including 16 SNPs from four independent loci based on single marker association, and 23 SNPs based on integration of gene association ranks from different data types and protein complex ranks based on pathway analysis.

### Validation

Based on the findings from the marker selection approaches described above (TDS single marker association, integration of gene-phenotype association and pathway analysis) and previously reported *KITLG* markers with association to TGCC, we selected a total of 39 SNPs (Supplementary Table 2) for subsequent validation in an independent replication cohort of 671 Nordic men. Out of the 671 samples, 23 did not meet the quality control standards for the genotyping system used by KBio-sciences.

Markers at three genes showed a significant association in the replication cohort; *KITLG*, *TGFBR3* and *BMP7* with odds ratios (ORs): 1.93 (1.39-2.69), 1.52 (1.08-2.15), 1.28 (1.01-1.62)) (Table 7.1). Surprisingly, the nine SNPs located in the *HOXD*





**Figure 7.3** – Regional association plots and linkage disequilibrium structure (LD). The  $-\log_{10}$  of the  $P$  value for the association of discovery markers are shown, and the markers are coloured in a white to red scale according to the strength of the pairwise LD ( $r^2$ ) to the most significant discovery marker at each locus. The association in the replication stage is represented by a blue marker. Light blue peaks indicate recombination rates from the CEU Hapmap population.  $r^2$  based LD structure from the GWA data are displayed at the bottom. (a) The strongest association in the discovery stage were found at 2q31.1, close to the cluster of *HOXD* genes. (b) The most significant markers in the replication stage were found at the *KITLG* region. The genes (c) *TGFBR3* and (d) *BMP7* had SNPs with mild association in the discovery stage but were top-ranked by the integrative data analysis, and were further validated in the replication stage.

region that were identified by the single marker association in the discovery cohort, did not find support in the validation cohort (Table 7.1 and Figure 7.3). Comparison of ORs for the different phenotypes of TDS, demonstrates that *KITLG* is mainly associated with TGCC but not with the other sub-phenotypes, whereas *TGFBR3* and *BMP7* show similar association for all phenotypes of TDS (Table 7.2).

## Discussion

Disorders of male reproduction have gained attention due to trends of increasing incidence and a growing health care burden in several countries. This study fo-

Gene	Marker	Phase	RA	TDS	TGCC	Odds Ratio		
						Cryptorchidism	Hypospadias	Infertile
<i>KITLG</i>	rs1352947	Discovery	T	1.52 (1.18-1.97)	2.74 (1.82-4.26)	1.08 (0.77-1.55)	1.00 (0.52-2.11)	1.26 (0.86-1.90)
		Replication		1.93 (1.39-2.69)	2.23 (1.55-3.28)	1.38 (0.89-2.24)	na	na
<i>TGFBR3</i>	rs12082710	Discovery	T	1.77 (1.33-2.36)	1.55 (1.07-2.23)	1.62 (1.07-2.44)	2.04 (0.90-4.56)	2.38 (1.54-3.70)
		Replication		1.52 (1.08-2.15)	1.49 (1.04-2.14)	1.64 (0.99-2.71)	na	na
<i>BMP7</i>	rs388286	Discovery	C	1.36 (1.11-1.67)	1.38 (1.06-1.79)	1.49 (0.78-1.52)	1.24 (0.70-2.22)	1.21 (0.89-1.66)
		Replication		1.28 (1.01-1.62)	1.37 (1.06-1.78)	1.09 (0.78-1.52)	na	na
<i>HOXDx</i>	rs17198432	Discovery	A	2.58 (1.82-3.70)	2.85 (1.86-4.38)	2.90 (1.81-4.63)	2.83 (1.11-6.64)	1.76 (1.01-3.01)
		Replication		0.97 (0.64-1.48)	0.96 (0.61-1.50)	1.00 (0.52-1.83)	na	na

**Table 7.2** – Associations with TDS and its sub-phenotypes, TGCC, cryptorchidism, hypospadias and infertile for replicated markers and the top marker in the discovery stage. na: not available, as these phenotypes were not assessed in the replication stage.

cused on understanding the genetics behind the developmental TDS, which is typically associated with male infertility and TGCC, sometimes combined with mild forms of genital malformations. Some studies suggest that a common genetic background for TDS is doubtful<sup>1</sup> and others even implied that TDS does not exist.<sup>27</sup> We strongly believe that TDS is a real syndrome with predominantly environment-dependent pathogenesis and a relatively weak component of genetic predisposition. That means that the prevalence of TDS among Caucasian populations most likely depends on the strength of the adverse environmental or lifestyle factors. However, as this study documents, it is likely that a subset of cases afflicted by TDS do share predisposing genetic variants. Here, we identified an association between several TDS phenotypes, most notably cryptorchidism and testicular cancer, and genomic variants in *TGFBR3* and *BMP7*. We also found evidence for an association with *HOXD* in the discovery cohort, but this was not confirmed in the validation study. *TGFBR3* has not previously been reported to be directly associated with male infertility or TDS, but the pathways this gene is involved in are highly relevant in terms of their involvement in regulation of embryogenesis and oncogenesis.

The most significant univariate association to TDS in the discovery cohort was rs17198432 and neighbouring SNPs, located at lunapark (*LNP*) and the *HOXD* cluster of genes, which were associated with several of the TDS phenotypes (Table 7.2, Supplementary Table 7). *HOXD* genes encode a group of transcription factors well known for their importance in morphogenesis during early embryonic development of limbs and genitalia<sup>28</sup> and a direct association with cryptorchidism has been shown in one study.<sup>29</sup> Recent functional studies of *HOXD* genes have shown the existence of a cisacting regulatory region located centromeric to *LNP*, a genomic control region (GCR) termed "limb enhancer", that modifies the expression of *HOXD* genes in early embryonic stages.<sup>22,30</sup> Both the GCR and the nine significant markers detected in this study are localized centromeric to the *HOXD* cluster and *LNP*, and may well lie within hitherto undiscovered regulatory elements. The mechanisms of *HOXD* regulation are not completely understood, but there exists some evidence for long range regulation: it has for instance been shown that large genomic deletions that include *LNP* lead to more severe defects than small deletions of *HOXD* alone.<sup>31</sup> *HOXD* was not supported in the replication cohort, but the variant may have low penetrance, and it cannot be ruled out that this discovery could be

validated in other cohorts. Notably, we observed that the risk allele for the marker with the highest significance at this locus, rs17198432, is very rare or even absent in African and Asian populations (Supplementary Table 8) which is concordant with the geographical patterns of TDS and testicular cancer incidence.<sup>3</sup>

Our study suggests that the genetic background for testicular cancer may be stronger than that of other TDS phenotypes. The analysis of the TGCC subset of cases confirmed the previously published association of the *KITLG* locus (12q21) with increased risk of testicular cancer, but without association for other subtypes of TDS. As discussed in detail earlier,<sup>13,14</sup> the relationship to TGCC is significant from a biological point of view. We believe that this finding does not necessarily characterize TDS, but only the testicular cancer sub-phenotype. However, it has been suggested that SNPs within the *KIT* and *KITLG* genes also may be involved in infertility.<sup>32</sup> Further, fine-mapping of the *KITLG* region remains to be performed to pinpoint causative variants within *KITLG* or its regulatory regions.

One aim of this study was to determine pleiotropic genetic risk factors that are associated with TDS as a whole, and thereby also differentiating associations that are specific to TGCC. Another objective was to identify functional variants that are undetectable by single marker association. This was done by application of systems biology methods that augment the available GWAS data with prior biological knowledge. We implemented two approaches, a pathway based analysis to identify protein complexes enriched for associated markers, and a gene-scoring method that combines multiple types of data, each of which provide estimates for the strength of association between genes and phenotypes.

The integrative systems biology analysis which combined data from the GWAS with gene-phenotype associations found in complementary data types, including testis developmental expression data and protein-protein complexes associated with testicular developmental defects, identified several disparate loci that are functionally linked to the TGF $\beta$  superfamily signaling pathway, which were supported by the replication study. Using GWAS alone, these SNPs would fall into a “gray zone”,<sup>17</sup> and would not have been selected for replication. Many GWA studies suffer from this phenomenon because the number of samples required to attain genome-wide significance can be unfeasibly high. Among the loci with the highest significance across all four TDS phenotypes was the *TGFBR3* gene, which encodes the TGF $\beta$  receptor type III, a co-receptor for inhibins and TGF $\beta$  1-3, BMP2, BMP4 and BMP7. TGFBR3 and its co-receptors and ligands are expressed in most endocrine tissues, including the testis. TGFBR3 has been identified in human Sertoli cells, both in the normal testis and in seminiferous tubules with CIS.<sup>33</sup> Importantly, TGFBR3 is essential for embryonic development of the reproductive system, as silencing of the murine *Tgfr3* resulted in impaired function of fetal Leydig cells and testicular dysgenesis.<sup>34</sup> Additionally, the signaling partners of TGFBR3, the family of activin receptors, are present in the fetal human testis<sup>35</sup> and appear to be dysregulated in testes with TGCC.<sup>36,37</sup> Of note, among the loci that scored high in both the pathway analysis as well as the integration analysis, were BMP7 and other bone morphogenetic proteins, which can bind TGFBR3. The presence of several components of the same pathway, which all have previously been implicated in the development of the testis and early reproductive system, reinforces the validity of

the identified SNPs as possible genetic factors predisposing to TDS. Dissecting the molecular consequences of the SNP variants of *TGFBR3* and *BMP7* genes will be a formidable task, because the biological function of these molecules in the testis has not been elucidated.

Although the candidate genes resulting from previous and present GWAS have biological functions in several adult tissues, they play particularly crucial roles in early differentiation and development of the gonads. Accumulating scientific literature suggests that a broad spectrum of genetic aberrations can lead to TDS, which in the most serious phenotypes can be classified as a disorder of sex differentiation.<sup>38</sup> These aberrations include, among others, aneuploidy of the sex chromosomes and chromosome 21 (Down's syndrome), mutations in sex determining genes e.g. *SRY*, *AR* (the androgen receptor) or *NR5A1* (*SF1*). These genetic aberrations lead to different phenotypes, however, at the gonadal level, all of them are associated with dysgenesis of the Sertoli and Leydig cells, which are assumed to be the primary targets in TDS and whose dysfunction may lead to failure of normal differentiation of germ cells and TGCC. In this context it is interesting that *KITLG*, a candidate gene of previous studies,<sup>13,14</sup> which was also validated as a TGCC gene in our study, is physiologically expressed in Sertoli cells. We speculate that abnormal expression of *KITLG* in Sertoli cells due to mutations, polymorphisms, exogenous exposures or androgen insufficiency may play a role in the early initiation of CIS from KIT-positive gonocytes. Similarly, inappropriate *TGFBR3* and *BMP* expression in the fetal gonad may contribute to testicular dysgenesis.

## Conclusion

Here, we successfully applied systems biology methodologies to prioritize important findings in a GWAS of a carefully collected discovery cohort of individuals from Danish ethnic background with detailed clinical records, and a subsequent replication cohort with similar phenotypes but possible genetic differences. The GWAS data was complemented by integration of multiple data types. This approach enabled us to identify potentially significant variations that would normally not be prioritized based on genome-wide multiple testing corrected p-values. Two out of the three replicated loci were selected based on systems biology, indicating that integrative analyses can be useful for candidate selection of markers that do not reach genome-wide significance in a single marker association analysis. The candidates identified in this study, notably those involved in  $TGF\beta$  superfamily signaling pathways, provide evidence that at least a subset of TDS cases may share common genetic predisposition. The paucity of informative markers indicates a predominant role of environmental and life style factors in the pathogenesis of this syndrome, at least in the Caucasian population. The findings will be corroborated in further independent GWAS, fine-mapping and sequencing studies, or by mechanistic investigations of the implicated pathways. Nevertheless, integration of prevailing information via systems biology approaches holds the promise to enhance future GWAS.

## Materials and methods

### Study subjects

All study subjects provided informed consent with approval of the local ethical committees. The discovery cohort constituted 926 individuals of Danish descent, 439 healthy young men with semen concentrations above 60 million sperm/ml (Table 7.3), and 488 cases affected by at least one of four TDS phenotypes, i.e. testicular germ cell cancer, cryptorchidism, hypospadias, or infertility. Specifically, the case group included 212 men with germ cell tumors, 138 men with cryptorchidism (diagnosed with lack of testis descent at birth), 31 men with hypospadias and 107 infertile men identified using the following criteria: absence of any known cause of infertility (no Y-chromosome microdeletions and no history of radio or chemotherapy) absence of varicocele, sperm count below 15 million sperm/ml, and testis volume below 15 ml. Cases with several symptoms of TDS were grouped according to severity (TGCC>hypospadias>cryptorchidism>infertility) (Supplementary Figure 1 and Supplementary Table 1).

For the replication phase, we obtained a total of 235 controls and 436 TDS cases from Denmark, Sweden and Finland: cases with cryptorchidism (n=103) were contributed from Finland (Turku area) (n=34) and from Rigshospitalet Denmark (n=69); TGCC samples (n=333) were collected in Sweden (from the Malmö region (n=112) and the Stockholm area (n=135)) and Denmark (n=86); controls were obtained from Rigshospitalet, Denmark (n=184) and the Turku area in Finland (n=51) (Table 7.3).

Phenotype	Discovery cohort	Replication cohort
Controls	439	235
Infertile	107	0
TGCC	212	333
Hypospadias	31	0
Cryptorchidism	138	103

Table 7.3 – Discovery and replication cohort constitution.

### Genotyping

Genomic DNA from peripheral blood was isolated from all cases and controls of the discovery cohort and genotyped using Affymetrix’s Genome-Wide Human SNP Array 6.0, following the manufacturer-supplied protocol. Genotype calling and subsequent analysis was performed using the Birdseed algorithm (of Affymetrix Power Tools), Plink v1.06 and R. Genotyping of 39 associated markers (Supplementary Table 2), selected from the discovery stage, was performed using the KBiosciences Competitive AlleleSpecific PCR SNP genotyping system (KASPar) as an in-house service at KBioscience. KASPar assay primers were designed using the Primer Picker software available at <http://www.kbioscience.co.uk/primer-picker.htm> (KBiosciences). Statistical analysis for single marker association was performed using R and plink (v. 1.06). P-values were obtained by the MAX test. Odds ratios with confidence limits were retrieved by logistic regression in R.

### Quality control

From the 926 samples included in the initial discovery cohort, we removed 55 samples where genotyping call rates were below 96% or their QC contrasts were below 0.4. Individuals detected to be of non-Scandinavian ancestry (n=11) were excluded (Supplementary Figure 3). In addition, 21 further samples with high potential of confounding the association analysis due to risk of consanguinity (n=17) or a high degree of missing data (n=4) were excluded.

### Single marker association statistics

We removed markers with low minor allele frequencies (MAF < 0.1), excess of missing genotypes (> 0.01) or strong deviation from Hardy-Weinberg equilibrium (< 10e-5). The resulting 600,798 markers were tested for association with the disease phenotypes using the MAX statistics.<sup>21</sup> Multiple testing correction was performed using the maxT permutation method.

### Gene-level meta-rank analysis

Metaranker<sup>19</sup> was used to integrate prior phenotype-specific data sets with the GWAS data to improve detection power. In short, the methodology can be subdivided into two steps: First, each data type is analyzed separately to produce a ranked list of all human genes based on the gene-phenotype association of that specific data type. Second, the ranked gene lists of each data set are combined into a single meta-rank of all human genes where each layer contributes equally to yield a final list of candidate susceptibility genes.

Data type specific gene lists were generated by analyzing four different data types to produce three individual gene rankings: (1) the genome-wide association data presented here; (2) targeted mutations resulting in testis developmental effects from the Mouse Genome Informatics (MGI), which were integrated with protein-protein interaction data;<sup>26</sup> (3) microarray gene expression data from the developing fetal testis of mouse<sup>23,24</sup> and human.<sup>25</sup> In the following we describe how each individual data layer was constructed.

#### Layer 1: GWAS evidence

To rank genes based on the GWAS we mapped markers to genes using Affymetrix Genome-Wide Human SNP Array 6.0 annotations. We assigned per-gene p-values by selecting the minimum p-value among the set of markers mapping to a gene. To avoid biases due to gene size, linkage disequilibrium and marker density, we corrected the pergene p-values by the number of independent markers.<sup>39</sup> The minimal p-value among all markers mapping to a gene was corrected by a Sidak multiple testing adjustment<sup>40</sup> using the latent number of independent markers. These corrected per-gene p-values entail a ranking of all human genes based on significance of association from the TDS GWAS.

### Layer 2: Expression data from fetal testis

To identify genes that are differentially expressed in the developing fetal testis we analyzed three array time-series expression data sets from mouse and human fetal testis, which were retrieved from the Gene Expression Omnibus (GEO) and normalized by *qspline* normalization.<sup>41</sup> The only human study of fetal testis found in GEO constituted 17 testis samples taken at time points between 9 and 20 weeks of gestation.<sup>25</sup> Data from mouse fetal testis included a study with biological duplicates at five time points, from the time of the indifferent gonad (11.5dpc) to birth (18.5dpc)<sup>23</sup> and a study of five time points ranging from gestational day 11 to 18.<sup>24</sup> For each mouse study, all genes were ranked based on their differential expression during testis development by grouping the samples by time-point and performing one-way analysis of variance (ANOVA) F-tests. Since the human study did not involve replicates, genes were instead ranked by the coefficient of variation (the standard deviation divided by the mean), which can be viewed as a generalization of fold-change to several groups. To generate one single ranked list of genes that describe the transcriptional regulation in the developing testis, the three separate rankings were combined into one rank by first mapping the mouse genes to their human orthologs and then using the sum of the individual ranks from the three studies.

### Layer 3: Mouse targeted genes with protein-protein complexes

All targeted alleles with their human orthologs and associated morphologies were obtained from MGI BioMart. A set of 34 morphologies from the Mammalian Phenotype ontology, were manually curated based on their developmental defects of the testis and relation to TDS (Supplementary Table 5). To avoid a bias towards spermatogenesis related genes, two morphologies, "male infertility" and "abnormal spermatogenesis", were excluded. All descendants to the curated terms were then retrieved from the Mammalian Phenotype ontology, resulting in a total of 122 terms. This set of terms was mapped to 315 human orthologous genes that were targeted in MGI. Next, we ranked all human protein complexes according to their enrichment for these 315 targeted genes. A complex was defined as a central hub protein and all its first-order interaction partners in a network of protein-protein interactions.<sup>26</sup> Each complex was tested for enrichment of targeted genes using a hyper-geometric test. In this manner, all sets of genes interacting in a biologically active complex obtained a rank with respect to its enrichment of TDS targeted mouse genes.

### Integration of three evidence layers into a meta-rank

We combined the single data-type evidence layers described above into a single metarank. First, all human protein-coding genes were ranked based on their significance in each of the three layers separately. This resulted in a rank,  $r^{L_i}(g)$ , for each gene,  $g$ , in every layer, where  $L_i$  denotes one of the layers in the set  $L = \{L_1, L_2, L_3\}$ . We then calculated a meta-score  $s_{meta}(g)$  for each gene as the

sum of its ranks in the three layers:  $s_{meta}(g) = \sum_{L_i \in L} r^{L_i}(g)$ . This meta-score was used to generate the meta-ranking of all human genes, where genes at the top of the list represent candidate disease genes.

### Pathway-based analysis of protein complex association

In order to rank protein complexes based on their likelihood to contribute to the TDS etiology (as opposed to single SNPs or single genes), we first generated a list of all human protein complexes using a network of human protein-protein interaction (PPI) data.<sup>26</sup> We defined a protein complex as a gene product and all its direct protein interaction partners. By iteratively selecting each gene together with all its interaction partners from the PPI network, we compiled a comprehensive list of protein complexes representing the complete human interactome up to the degree for which experimental data is available. We merged complexes with pairwise similarity above 80% using the Jaccard index as a metric. As described above for the creation of a ranked gene list from the GWAS data, each gene in a protein complex was assigned a p-value based on the minimal p-value of all markers mapped to that gene, and corrected for gene size and LD biases by a Sidak multiple testing correction using the number of independent markers of a gene. This correction effectively eliminates a bias of larger genes systematically receiving low p-values due to a higher coverage of SNPs.<sup>39</sup> Additionally, in a small number of cases we removed genes from a complex if they were co-localized in blocks of high linkage disequilibrium, and complexes with more than 23 genes were removed since those complexes often are largely overlapping with many smaller complexes, and their biological interpretation is unclear. The resulting 8,342 complexes were tested for enrichment of genes with low p-values and ranked accordingly. By combining the per-gene p-values for each complex into a per-complex score using a Z-score transformation<sup>42</sup> the resulting 8,342 complexes were tested for enrichment of genes with low p-values and ranked accordingly. A final ranking was established by comparing the per-complex scores against a sampled background distribution of randomly generated complex-scores. We manually selected complexes from the top of the list and were specifically interested in complexes that contained genes identified by the meta-ranking analysis described above.

### Acknowledgements

We would like to thank Betina F. Nielsen for her excellent and careful microarray work. We thank all patients for participation in this study, and the health professionals for facilitating our work. We would like to acknowledge support from the VillumKann Rasmussen Foundation, a NABIIT grant from the Danish Strategic Research Council, and the Novo Nordisk Foundation, the Academy of Finland, Sigrid Juselius Foundation, Foundation for Paediatric Research, Turku University Hospital, and The European Commission (FP7/2008-2012: DEER 212844). The Swedish Cancer Society (CAN 2009/817), Gunnar Nissons Cancer Foundation, Malmö University Hospitals Cancer Foundation and King Gustaf V's Jubilee fund for Cancer Research



### **Author contributions**

MDD, HL, NJ, AJ, SB, RG, TSJ, NES and ERM designed the study; MDD, NW and DE performed experiments; NW, DE, MDD, JDS, TAG, THP, TSJ and RG collected and analyzed data; NJ, ERM, AG, YLG, GCC, HEV, JT and GD provided DNA samples and materials; NW, DE, MDD, HL, ERM, NES, SB and RG wrote the manuscript; SB, AJ, TSJ and THP gave technical support and conceptual advice.

### **7.3 Supplementary material**

Available in Appendix B.

---

## Chapter 8

# Rare copy number variations affecting the risk for testicular germ cell tumor

---

### 8.1 Abstract

Testicular germ cell tumor (TGCT) is one of the most heritable forms of cancer. Previous genome-wide association studies have focused on single nucleotide polymorphisms (SNPs), largely ignoring the influence of copy number variants (CNVs), which are poorly understood. Here we present the results from a whole-genome CNV study on a cohort of 212 cases and 437 controls from Denmark, who were genotyped at 1.8 million markers, half of which were non-polymorphic copy number markers. We observed a higher accumulation of rare CNVs (present in no more than 1% of the samples) in cases as compared to controls, at the gene *PTPN1* ( $P=3.8 \times 10^{-2}$ , 0.9% of cases and 0% of controls), as well as for genes annotated as regulators of cell migration (FDR=0.021, 1.8% of cases and 1.1% of controls). In independent control cohorts of healthy individuals no CNVs were observed at *PTPN1*. The *PTPN1* phosphatase is involved in key signaling pathways and has previously been implicated in multiple cancers. Dysregulation during migration of primordial germ cells has previously been suspected to be part of TGCT development and we demonstrate support for such an etiology. Our findings indicate that multiple rare variants contribute to the pathogenesis of TGCTs and implicate several new genes that may cause increased susceptibility.

## 8.2 Manuscript

### Rare copy number variations affecting the risk for testicular germ cell tumor

Daniel Edsgård D<sup>1,#</sup>, Marlene D. Dalgaard<sup>2</sup>, Nils Weinhold<sup>1</sup>, Ewa Rajpert-De Meyts<sup>2</sup>, Anne Marie Ottesen<sup>2</sup>, Anders Juul<sup>2</sup>, Niels E. Skakkebæk<sup>2</sup>, Thomas Skøt Jensen<sup>1</sup>, Ramneek Gupta<sup>1</sup>, Henrik Leffers<sup>2</sup>, Søren Brunak<sup>1</sup>

<sup>1</sup> Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

<sup>2</sup> Department of Growth and Reproduction, Rigshospitalet, Copenhagen, Denmark.

# Correspondence should be addressed to edsgard@cbs.dtu.dk

#### Author summary

Testicular germ cell tumor is the most common cancer among young men. Although treatments for this type of cancer are exceptionally effective (up to 95% of patients can be relieved from cancerous symptoms), the cancer and its treatment options are often accompanied by infertility problems. The disease is known to both have a genetic cause, evident from familial aggregation, as well as being related to environmental exposures. No single defective gene has been found to explain a large proportion of cases, and the genetic contribution to the disease is therefore likely to originate from multiple genetic variants. In this study we identified copy number variants within the genome that predispose to the disease by examining the DNA from 650 Danish individuals. Comparing affected cases to healthy controls, we find an increased number of rare copy number variants at the gene *PTPN1* among cases, and at genes related to cell migration. These findings will help to pinpoint biological pathways and genetic variants that are of particular importance in the pathobiology of testicular germ cell tumors.

#### Introduction

Testicular germ cell tumor (TGCT) is the most common malignancy in young men aged 15–45 years. The incidence has increased over the last decades and is highest in Nordic countries with 8–9 cases per 100,000, whereas the incidence in men of African and Asian ancestry is five-fold lower.<sup>1</sup> Environmental exposure partly explains the increasing incidence and ancestral disparity, but there is also evidence of a substantial genetic contribution to TGCT susceptibility. The familial aggregation of TGCT is one of the highest among cancers. Brothers and sons of TGCT patients have a 8–10 times and 4–6 times higher risk to develop the disease, while the risk increases 75- and 35-fold for monozygotic and dizygotic twins, respectively.<sup>2,3</sup> Despite the relatively high degree of heritability, genome-wide familial linkage analyses have not identified any loci predisposing for TGCT, and candidate studies

have only found one rare deletion (2-3%) at the Y chromosome that confer a modest 2-3 fold increased risk.<sup>4</sup> Recently, genome-wide association studies that search for common single nucleotide variants associated to TGCT have identified susceptibility loci at the genes *KITLG*, *SPRY4*, *BAK1*, *DMRT1*, *TERT* and *ATF7IP*.<sup>5-7</sup> The strongest association was found at *KITLG* with a greater than 2.5-fold increased risk of disease. Consistent with the relatively high familial relative risk, this is the largest effect size found for any single loci among cancers. However, a considerable portion of the heritability remains to be explained.

Here we investigate constitutional DNA CNVs as another source of genetic variability that may contribute to the development of TGCT. Recent studies have described associations of common CNVs with neuroblastomas,<sup>8</sup> systemic autoimmunity,<sup>9</sup> psoriasis<sup>10</sup> and osteoporosis.<sup>11</sup> Rare variants, typically originating from recent and *de novo* events, constitute a significant portion of genomic variation. The thousand genomes project indicates that there are about twenty thousand CNVs with allele frequencies down to 1%.<sup>12</sup> The contribution of such rare, or even rarer, variants, to complex disease susceptibility is to a large extent unknown, but they seem to play an important role in psychiatric disorders<sup>13,14</sup> and they have been indicated to influence childhood obesity.<sup>15</sup> To date, association studies of individual rare CNVs have insufficient power to identify disease-causing variants. To evaluate the impact of rare CNVs with respect to risk for TGCT, we therefore compared the genome-wide burden of rare CNVs and investigated whether any genes or pathways were targeted by multiple rare CNVs such that their aggregated frequency was higher in cases than in controls.

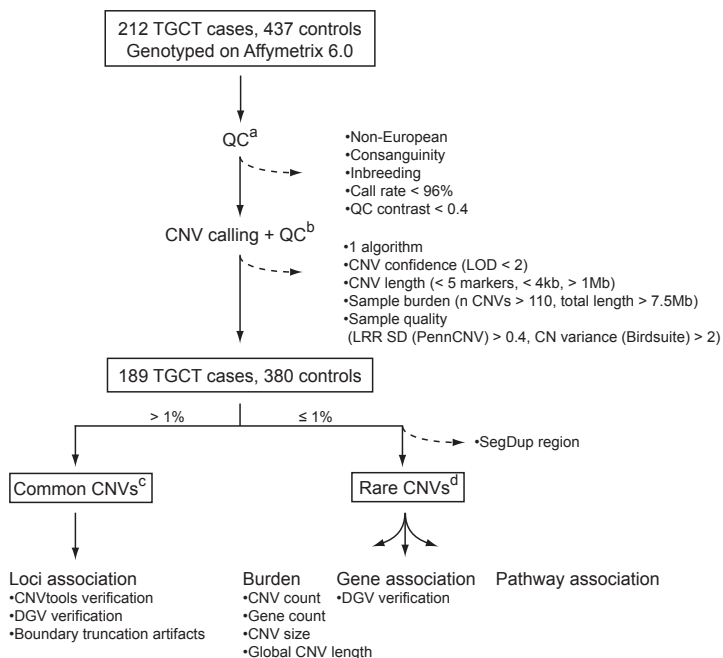
In summary, to assess the effect of copy number variants on TGCT we genotyped a Danish case-control cohort, as described elsewhere (Dalgaard *et al.*, submitted), and identified one potential gene and one protein network of particular interest, having a significantly higher frequency of rare CNVs among cases than among controls.

## Results

To identify CNVs that confer a risk to testicular germ cell cancer, we analyzed common and rare variants in a genome-wide dataset of approximately 1.8 million markers interrogating a Danish cohort constituting 212 TGCT cases and 437 controls (SNP association analysis and details of the cohort are described elsewhere, Dalgaard *et al.*, submitted). Application of stringent quality control criteria for reliable CNV identification (Figure 1, Methods) resulted in a final discovery set of 189 cases and 380 controls. Common variants were defined as CNVs present in more than 1% of the study population, and rare variants as CNVs present in no more than 1% of the studied subjects. Common variants were analyzed with respect to individual locus association, and rare variants with respect to overall genetic burden, gene association and pathway association.

### Locus association analysis

In order to identify common CNVs associated with TGCT, binary copy number state frequencies of the case and control cohorts were compared at all loci with CNV



**Figure 8.1** – CNV quality control and analysis. Dashed arrows indicate CNVs and samples that were excluded from the analysis: *a*) quality control of samples based on SNP calls; *b*) quality control of CNVs and samples based on CNV calls; *c*) association analysis of common CNVs; *d*) association analysis of rare CNVs with respect to genomic burden, as well as genes and pathways with an excess of rare CNVs among cases. See Methods for further details. LOD, log odds; LRR SD, log R ratio standard deviation; DGV, Database of Genomic Variants.

frequencies above 1%. We observed one genome-wide significant deletion at 1p13.3 covering the gene *GSTM1* ( $P = 0.02$ , 37.2% cases, 19.5% controls), but downstream quality control by manual inspection of the copy number intensity histogram at this locus, and application of histogram-based association analysis<sup>16</sup> suggested a false positive finding (nominal  $P = 0.07$ , 48.7% cases, 41.8% controls, which was further supported by a deletion allele frequency at this locus estimated to be 40% in the International HapMap Phase 3 population study.<sup>17</sup>

### CNV burden association analysis

Testing whether individuals with TGCT had a greater genomic burden of rare CNVs than controls, we observed a weak indication of increased burden with respect to the number of CNVs per sample, the number of affected genes per sample, and the average length of CNVs per sample (case/control ratio: 1.08, 1.10 and 1.11 respectively), and a significant difference with respect to the total length of all CNVs per sample (case/control ratio: 1.19,  $P = 0.03$ ) (Table 1).

Type	Burden	<i>P</i>	Case/control ratio	Baseline (ctrl)	Baseline (case)
All	Rate	0.09	1.08	4.8	5.17
	Gene rate	0.19	1.10	3.3	3.7
	Mean length (kb)	0.14	1.11	76.9	85.1
	Total length (kb)*	0.03	1.19	372.5	444.3
Duplications only	Rate	0.36	1.03	1.9	1.93
	Gene rate	0.29	1.09	1.9	2.1
	Mean length (kb)*	0.01	1.33	112.3	149.0
	Total length (kb)	0.07	1.22	279.4	340.8
Deletions only	Rate	0.09	1.12	2.9	3.24
	Gene rate	0.23	1.11	1.4	1.6
	Mean length (kb)	0.49	1.00	49.9	49.9
	Total length (kb)	0.21	1.09	151.6	165.0

**Table 8.1** – Global burden of rare CNVs in cases versus controls. The table shows the rate and length of CNVs in cases versus controls. Genome-wide *P*-values were estimated by 10,000 permutations of case-control status. \*Significant difference ( $P < 0.05$ )

### Gene association analysis

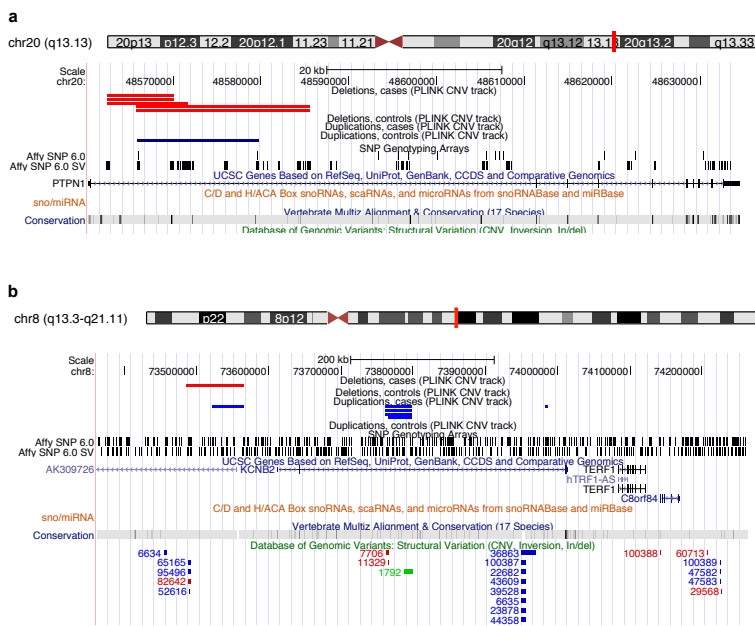
Next, we explored if there were any specific genes where rare CNVs were more common in cases than controls. This analysis did not require that CNVs found in different samples overlapped each other, rather, it was sufficient that they were located within the same genic region. Two genes were found to have genome-wide significance, *PTPN1* ( $P_{emp.} = 0.038$ ) and *KCNB2* ( $P_{emp.} = 0.022$ ), affecting 0.9% and 1.2% of cases, respectively, whereas no occurrence in controls was observed at these loci (Table 2). The CNV at *PTPN1* involved five cases, all found to have a heterozygous deletion at the same intronic region (Figure 2). CNVs at *KCNB2* were found at three different loci: four and one deletions at two different introns; and one deletion and one duplication at the promoter (Figure 2). Several CNVs have previously been reported in healthy individuals at *KCNB2*, but not at *PTPN1* (Database of Genomic Variants, v. 10), corroborating a true TGCT association at *PTPN1* but weakening the possibility of an actual association to *KCNB2*.

Gene	$P_{emp.}^1$	$P_{nom.}$	OR	Cases	Controls
<i>PTPN1</i>	$3.8 \times 10^{-2}$	$3.9 \times 10^{-3}$	12.31 (1.48-568.17)	0.9%	0.0%
<i>KCNB2</i>	$2.2 \times 10^{-2}$	$4.1 \times 10^{-4}$	16.58 (2.19-738.20)	1.2%	0.0%

**Table 8.2** – Genes with an association of rare CNVs. <sup>1</sup>Empirical genome-wide *P*-values were estimated by 1,000 permutations of case-control status using Fisher's test. nom.:nominal *P*-value.

### Pathway association analysis

Proteins tend to act in concert and perturbations of different components in a set of proteins that are interacting in a network may result in a dysregulation with similar outcome.<sup>18</sup> We therefore conducted association analysis on the level of pathways and protein-protein interaction networks. Association was assessed in the



**Figure 8.2 – Genes with a significant excess of rare CNVs among cases. (a)** Five cases with deletions and one control with a duplication at an intron of *PTPN1*, **(b)** Six cases with deletions and one case with a duplication at introns of *KCNB2*. The bottom track (Database of Genomic Variants) indicates that many CNVs have previously been observed at *KCNB2* in healthy individuals, whereas none has been observed at *PTPN1*.

same fashion as for genes and loci described above, that is, by comparing case and control cohorts with respect to the total frequency of rare CNV events targeting a pathway. We compiled comprehensive collections of gene sets, where a set of genes either share function or operate in the same pathway, and performed one thousand random permutations (shuffling case and control status) to estimate a local false-discovery rate (FDR) of each gene set.<sup>19</sup> We observed a significant differential proportion of rare CNVs between cases and controls in 14 gene sets (FDR < 5%, Table 3). However, eleven of these associations were driven by the gene-specific association to *PTPN1* described above. The three remaining gene sets, which did not include *PTPN1*, were: ‘regulation of cell migration’ (go:0030334, 1.8% versus 1.1%, FDR = 0.021, OR = 3.47 (1.12–11.82)), ‘positive regulation of catalytic activity’ (go:0043085, 1.4% versus 0.5%, FDR = 0.04, OR = 5.54 (1.31–32.78)), and ‘macromolecular complex disassembly’ (go:0032984, 2.3% versus 1.9%, FDR = 0.047, OR = 2.47 (1.00–6.23)). The most significant gene set among all sets tested was ‘regulation of cell migration’. A total of 16 individuals harbored CNVs that altogether overlapped 14 genes in this pathway, at 13 unique CNV loci (Table 4). Genes in this gene set that were affected by cases but not controls included: *BCL2*, *CDH13*, *CORO1A*, *KDR*, *MUC2*, *MUC5AC*, *ONECUT2* and *PTPRK*.

Gene set type <sup>1</sup>	Term name	Posterior <sup>2</sup>	Local FDR	OR	Cases	Controls
GO BP	Regulation of cell migration	0.98	0.021	3.47 (1.12-11.82)	1.8%	1.1%
GO BP	Macromolecular complex disassembly	0.96	0.040	5.54 (1.31-32.78)	1.4%	0.5%
GO BP	Positive regulation of catalytic activity	0.95	0.047	2.47 (1.00-6.23)	2.3%	1.9%

**Table 8.3** – Gene sets with an association of rare CNVs. <sup>1</sup>Gene sets from many sources were jointly analyzed but only sets of the type ‘gene ontology biological process’ were significant, apart from gene sets that included *PTPN1* which were excluded from the table (see Supplementary Material). <sup>2</sup>The empirical Bayes analysis of microarrays (EBAM) algorithm with 1,000 permutations was used to estimate a posterior and local false discovery rate (FDR) for every gene set.

CNV	Length <sup>1</sup>	CN <sup>2</sup>	Sample ID	Class	Genes
chr4:55607652..55616597	9	1	165855	Case	<i>KDR</i>
chr6:128485528..128525520	40	1	232996	Case	<i>PTPRK</i>
chr6:128864684..128871092	6	1	190037	Case	<i>PTPRK</i>
chr11:1094626..1140711	46	3	210711	Case	<i>MUC2, MUC5AC</i>
chr16:29474810..30099408	625	1	124873	Case	<i>CORO1A</i>
chr16:29488112..30085920	598	1	233682	Case	<i>CORO1A</i>
chr16:82119367..82175095	56	1	224567	Case	<i>CDH13</i>
chr16:82408574..82502970	94	1	232030	Case	<i>CDH13</i>
chr18:52763504..53341297	578	3	203688	Case	<i>ONECUT2</i>
chr18:59018003..59031365	13	1	231734	Case	<i>BCL2</i>
chr2:55119289..56699138	1580	3	M3088A	Control	<i>RTN4</i>
chr3:188880305..188936673	56	3	M1270A	Control	<i>SST</i>
chr12:50532205..50579767	48	3	M3576A	Control	<i>ACVRL1</i>
chr15:50811752..50882082	70	3	M3047A	Control	<i>ONECUT1</i>
chr15:97371582..97730964	359	3	M053A	Control	<i>IGF1R</i>
chr19:49893892..50298979	405	3	M3381A	Control	<i>APOE</i>

**Table 8.4** – CNVs targeting genes that are part of the gene set ‘Regulation of cell migration’. <sup>1</sup>Kilobases, <sup>2</sup>Copy number.

## Discussion

We assessed the effect of common and rare copy number variants in a TGCT case-control cohort, and identified one potential gene and one network of particular interest, that had an elevated frequency of rare CNVs among cases. The absence of any common CNV associated to TGCT is in line with the relatively few findings for other diseases, including a recent screening of 3400 common CNVs in eight common diseases.<sup>20</sup> Furthermore, common CNVs are typically ancient variations, which are tightly correlated with SNPs, and can therefore be detected by genome-wide association studies of common SNPs.<sup>21</sup> However, one should not neglect the importance of common CNVs in gene-phenotype association studies, since there exists evidence that disease associated SNPs have a tendency to tag CNVs more often than random, and that such CNV-tagging SNPs are enriched for expression quantitative trait loci (eQTL).<sup>22</sup>

Here we report one gene in association with rare CNVs, *PTPN1*. A second gene, *KCNB2*, was considered a false positive due to the amount of reported CNVs in



independent control cohorts at this locus (DGV). *PTPN1*, encodes a classical non-transmembrane protein tyrosine phosphatase that plays a key role in signaling pathways. Several kinases have been described as substrates for *PTPN1*, including EGFR, JAK2 and TYK2, among others. These signaling networks are critical for cellular control, and *PTPN1* has been shown to be an important regulator involved in human disorders such as diabetes, obesity, and cancer.<sup>23</sup> Notably, CNVs at the 20q13 chromosomal region have been observed in several cancers<sup>24–26</sup> and associated with poor prognosis in breast cancer.<sup>27</sup> All affected probands in this study presented a deletion, consistent with a tumor suppressing function of oncogenic kinases. *PTPN1* has however been shown to be able to play both a pro- and anti-oncogenic role, probably reflecting an intricate dynamic signaling system and dependence on genetic background.<sup>28</sup> It is of interest that mutations in a related protein, *PTPN11*, are known to cause Noonan syndrome, a Mendelian disorder which causes abnormal development of multiple parts of the body, including short stature, delayed puberty, small penis and undescended testicles.<sup>29</sup> It is not unlikely that there is an overlap in substrate binding partners for *PTPN11* and those of *PTPN1*, as substrate sharing may occur among phosphatases.<sup>30</sup>

Our pathway analysis identified the gene set ‘regulation of cell migration’ as having the highest difference in proportion of rare CNVs in cases compared to controls. There is a growing body of information that strongly suggests a crucial role of primordial germ cell (PGC) biology in TGCT oncogenesis. PGCs are embryonic cells which during mid-gestation migrate from the base of the yolk sac, along the hind-gut, to the genital ridge, one of the longest migrations of all mammalian cells. Three recent SNP GWAS TGCT studies associated *KITLG* and a number of other genes related to the KIT-KITLG pathway,<sup>5–7</sup> a regulatory network which is believed to be of crucial importance in the determination of the fate of primordial germ cells.<sup>31</sup> For instance, mutations of the KIT receptor, or the KIT ligand, in the mouse, blocks PGC migration, resulting in infertility.<sup>32</sup> In addition, a disturbance of the migration of PGCs during early fetal development may cause extragonadal germ cell tumors along the midline of the body.<sup>33</sup> One of the afflicted genes in the ‘regulation of cell migration’ gene set was *PTPRK*, at which two samples had a deletion. Like *PTPN1*, *PTPRK* belongs to the family of protein tyrosine phosphatases, but it is a membrane-bound receptor. *PTPRK* has been implicated in TGFβ-signaling,<sup>34</sup> and a recent GWAS study indicated the involvement of TGFβ superfamily signaling in testicular dysgenesis (Dalgaard *et al.*, submitted). In mouse, PGCs divide rapidly under the influence of TGFβ signaling factors, and defects in PGC development is observed in knockout models of bone morphogenetic proteins (BMPs).<sup>35,36</sup>

In conclusion, this study suggests several rare copy number variants that contribute to the oncogenesis of a subset of TGCT. The frequency after aggregation of CNVs on the implicated gene and pathways is still low, and these CNVs therefore only provide a minor contribution to the overall heritability. Larger cohorts are needed to further explore the impact of rare variants.

## Materials and Methods

### Sample collection

212 men with testicular germ cell tumors and 439 healthy young men with semen concentrations above 60 million sperm/ml were collected at the Department of Growth and Reproduction, Rigshospitalet, Denmark. All samples were of Danish ancestry and provided informed consent (Danish ethical committee).

### Genotyping and SNP-based sample quality control

DNA from peripheral blood was extracted from all subjects and processed according to the manufacturer's protocol and samples were genotyped using the Affymetrix Genome-wide Human SNP Array 6.0. Here we present the analysis of CNVs, but an initial sample quality control was performed using SNP genotypes called by the Birdseed algorithm (Affymetrix Power Tools v. 1.10.2). We excluded samples based on a genotyping call rate below 96%; QC contrast below 0.4 as according to the Affymetrix GCOS software; non-European ancestry by inspection of a plot of the first two principal components of the cohort and the HapMap phase III samples;<sup>17</sup> high degree of relatedness based on identity-by-descent where one individual was kept among related samples; and samples with an outlying inbreeding coefficient.

### CNV detection and quality control

For samples that passed the SNP quality control described above we ran two CNV calling algorithms, BirdSuite (v. 1.5.5) and PennCNV (v. 2010May01). PennCNV requires a signal intensity file and a SNP genotyping file and these were generated by quantile normalization of PM-only probes with median polish probe summarization (Affymetrix Power Tools (APT) v. 1.12.0) and Birdseed (v. 2 in APT 1.12.0), respectively. We excluded CNVs which failed any of the four following criteria: (1) A CNV was not called by both algorithms. A histogram of the percentage of overlap indicates that the vast majority has >90% overlap, but we set the threshold to at least 10%. (2) A CNV log odds confidence score larger than two, as recommended by BirdSuite.<sup>37</sup> (3) CNV size was less than 5 markers or four kilobases, in effect excluding the 25% short-length quantile of CNVs. (4) A CNV was longer than one megabase (three outliers based on histogram). Further, samples were excluded with respect to the three following criteria: (1) extreme sample burden in terms of more than 110 CNVs (four outliers); or (2) a total length of CNVs larger than 7.5 megabases (two outliers); (3) bad sample quality in terms of high variance of copy number signal (median copy number variance larger than two, as recommended by BirdSuite, or a Log R Ratio standard deviation (LRR SD) obtained from PennCNV larger than 0.4). LRR SD was set according to PennCNV guidelines when CNV calling are performed on Affymetrix samples. Finally, rare CNVs that had more than 50% overlap with segmental duplication regions (retrieved from UCSC hg18) were removed, since such regions have been shown to generate more false CNV calls.<sup>14</sup> A total of 189 TGCT cases and 380 controls remained after the completion of all quality control steps, harboring a total of 1008 and 1872 rare CNVs, respectively.

### **CNV association analysis**

CNV association analysis was performed using plink (v. 1.07) and custom R (v. 2.12) scripts. Common and rare CNVs were defined as those with an allele frequency above or below 1%, respectively. The allele frequency of a CNV was determined using the locus within a CNV region with the maximum number of overlapping individual CNVs.

### **Locus association**

Common CNVs were evaluated by searching the whole genome for loci with a significantly higher degree of affected cases as compared to controls. Binary state frequencies were used, and a Fisher test was performed for deletion, amplification and any type of aberration, respectively. Genome-wide significance was estimated by generating a null distribution based on one thousand case-control status permutations. For each permutation the minimal  $P$ -value was selected, thereby providing control of the family-wise error rate (FWER). Loci with significant associations were further verified by CNVtools<sup>16</sup> which uses a complementary CNV-calling strategy, since it employs a statistical model based on density-based clustering rather than a hidden Markov model. Furthermore, loci residing at the edges of a common CNV and associations from variation of boundary truncation were excluded.

### **Global burden analysis**

The impact of rare CNVs was assessed by three approaches: genome-burden analysis, gene association and pathway association analysis.

The global burden of rare CNVs in cases compared to controls were assessed with respect to (i) the number of CNVs per sample, (ii) the number of affected genes per sample, (iii) the average length of CNVs per sample, and (iv) the total length of CNVs per sample.

### **Gene association**

Gene association analysis was performed using the number of case and control samples harboring a rare CNV that overlapped the gene of interest. Genes were retrieved from UCSC (hg18). CNV frequencies were compared with Fisher's test and multiple testing corrected  $P$ -values were obtained based on case-control permutation as described above for the locus association of common CNVs. Significant CNVs which were found to have a lower allele frequency in our case cohort than in the Database of Genomic Variants (DGV, v. 10), were considered false positives.

### **Pathway association**

Pathway association analysis was performed based on the number of case and control samples that had a rare CNV in any of the genes of a pathway. The R package 'siggenes' was used to obtain  $P$ -values corrected for multiple testing across all tested gene sets. The package provides a false discovery rate (FDR),

based on randomized case-control sampling, as well as an adjustment of the variance of an individual pathway using information from the observed variances of all pathways.<sup>19</sup> Gene sets were retrieved from KEGG (Kyoto Encyclopedia of Genes and Genomes),<sup>38</sup> Reactome,<sup>39</sup> BioCarta (<http://www.biocarta.com>), NCI-Nature curated pathways (Pathway Interaction Database),<sup>40</sup> GO (Gene Ontology),<sup>41</sup> COSMIC (Catalogue of Somatic Mutations In Cancer),<sup>42</sup> Cyclebase,<sup>43</sup> protein-protein interaction complexes,<sup>44</sup> OMIM (Online Mendelian Inheritance in Man, <http://www.ncbi.nlm.nih.gov/omim>), MGI (Mouse Genome Informatics, <http://www.informatics.jax.org>) and a set of candidate infertility genes from a recent review.<sup>45</sup> Terms annotating more than 700 or less than five genes were discarded, since they do not produce meaningful statistical results.

### Acknowledgements

We would like to thank Betina F. Nielsen for skillful microarray work and clinicians for handling of patients. The work was supported by a grant from the Villum Kann Rasmussen Foundation.

### 8.3 Supplementary material

Available in Appendix C.



---

## Chapter 9

# Next generation sequencing: RNA-seq of an unsequenced plant

---

### 9.1 Abstract

Studies of the resurrection plant *Craterostigma plantagineum* have revealed some of the mechanisms which these desiccation-tolerant plants use to survive environments with extreme dehydration and restricted seasonal water. Most resurrection plants are polyploid with large genomes, which has hindered efforts to obtain whole genome sequences and perform mutational analysis. However, the application of deep sequencing technologies to transcriptomics now permits large-scale analyses of gene expression patterns despite the lack of a reference genome. Here we use pyro-sequencing to characterize the transcriptomes of *C. plantagineum* leaves at four stages of dehydration and rehydration. This reveals that genes involved in several pathways, such as those required for vitamin K and thiamin biosynthesis, are tightly regulated at the level of gene expression. Our analysis also provides a comprehensive picture of the array of cellular responses controlled by gene expression that allow resurrection plants to survive desiccation.

## 9.2 Manuscript

### Transcriptomes of the desiccation tolerant resurrection plant *Craterostigma plantagineum*

Maria C. Suarez Rodriguez<sup>1,†</sup>, Daniel Edsgård<sup>2,†</sup>, Syed S. Hussain<sup>3</sup>, David Alquezar<sup>1</sup>, Morten Rasmussen<sup>1</sup>, Thomas Gilbert<sup>1</sup>, Bjørn H. Nielsen<sup>2</sup>, Dorothea Bartels<sup>3</sup>, John Mundy<sup>1,#</sup>

<sup>1</sup> Department of Biology, University of Copenhagen, Copenhagen, Denmark.

<sup>2</sup> Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark.

<sup>3</sup> Institute of Molecular Physiology and Biotechnology of Plants, University of Bonn, Bonn, Germany

† These authors contributed equally

# Correspondence should be addressed to [mundy@science.ku.dk](mailto:mundy@science.ku.dk)

### Introduction

Strategies for desiccation avoidance and tolerance are common in non-vascular plants and in the seeds, pollen and spores of tracheophytes. In vegetative tissues of higher plants, the ability to recover from extreme dehydration and to resume normal growth when water again becomes available is an exceptional characteristic of the resurrection plants.<sup>1</sup> Some 300 angiosperms, including a small number of dicotyledonous plants, exhibit such extreme desiccation tolerance and can be revived from an air-dried state.<sup>2,3</sup> A goal of research on these species is to discover genes and metabolic pathways that may help the design and production of desiccation-tolerant crops.

Previous studies have unveiled cellular mechanisms that contribute to desiccation tolerance. Less complex plants such as algae, moss and liverworts may rely more on overall cellular protection and repair to overcome sudden changes in water content.<sup>4</sup> Vascular resurrection plants display a constitutive system of protection against cellular damage that responds efficiently to gradual changes in water availability.<sup>5,6</sup> Two main groups of resurrection plants are distinguished by their capacity to degrade or retain their chlorophyll. *Craterostigma plantagineum* belongs to the homoiochlorophyllous resurrection plants that retain chlorophyll and thylakoid structures during dehydration. This requires specific metabolic regulation to preserve the structure of the photosynthetic apparatus and at the same time to provide efficient protection against the photo-oxidative properties of the remaining chlorophyll.<sup>7</sup>

*Craterostigma plantagineum* has become the preferred resurrection plant for molecular studies because it can be genetically transformed and it exhibits desiccation tolerance in both undifferentiated callus and in differentiated plants. Several

studies have elucidated specific features which allow *C. plantagineum* to tolerate desiccation to less than 5% relative water content (RWC). For example, a search for dominant mutants identified the unusual *CDT1* gene that is expressed upon dehydration and confers abscisic acid (ABA)-mediated desiccation tolerance to callus tissues, probably through a siRNA-mediated process.<sup>8,9</sup> Metabolic analyses found that the dominant carbohydrate in leaves is the rare sugar 2-octulose which is considered to be a reserve carbohydrate as it is rapidly converted to sucrose upon dehydration.<sup>10,11</sup>

Apart from these apparently unique features, the general dehydration responses in *C. plantagineum* and other resurrection plants are similar to those observed in non-tolerant plants. For example, several conserved sets of genes encoding proteins with diverse roles during water stress are tightly regulated. These include late embryogenesis abundant (LEA) proteins and molecular chaperones, enzymes required for carbohydrate metabolism and proteins that act to scavenge reactive oxygen species (ROS).<sup>5,12</sup> These intracellular proteins protect against water stress by stabilizing membranes and organelles, providing matrices for the integrity of molecular complexes, avoiding excessive loss of ions, and counteracting oxidative stress.<sup>13</sup>

To obtain a more complete picture of the genes and metabolic processes involved in the acquisition of desiccation tolerance, we undertook a comprehensive study of gene expression profiles in *C. plantagineum* at crucial stages of dehydration and rehydration. Since a *C. plantagineum* genome sequence was not available, deep expression profiling was only feasible by large-scale pyro-sequencing of transcripts from four hydrated and dehydrated *C. plantagineum* leaf samples. Assembly of roughly 182 Mb of transcript sequences yielded more than 15 000 significant UniProt identities. A gene ontology (GO) enrichment analysis on the 500 most variable transcripts across the samples revealed overrepresented biological functions and metabolic processes that provide hallmarks for the stages of hydration tested in *C. plantagineum*.

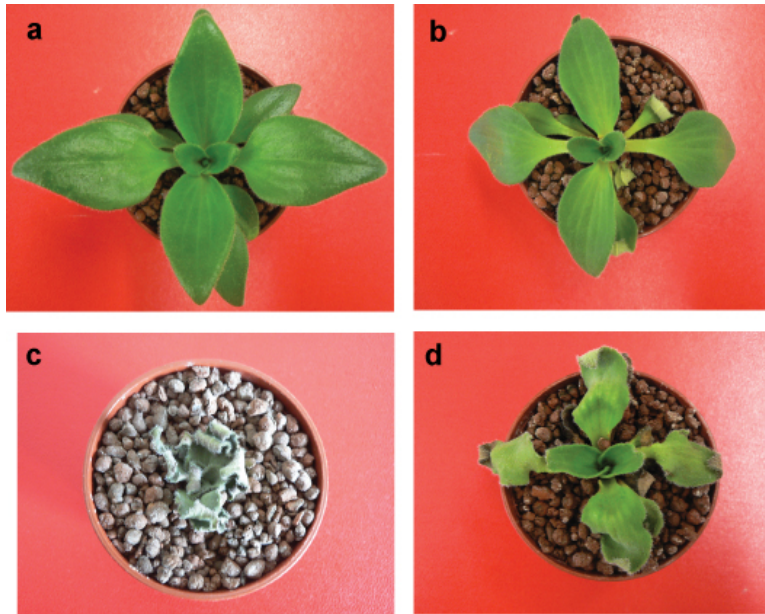
Comparison of our data with previously reported transcript profiles in orthodox seeds and pollen indicates that vegetative desiccation tolerance may be the result of differential regulation of pre-existing, non-vegetative desiccation mechanisms. In addition, comparison with transcript profiles of desiccation-sensitive plants indicates that most water-stress-related genes are shared by tolerant and nontolerant species, but that changes in their expression patterns ultimately provide tolerant plants with more effective protective mechanisms.

## Results

### EST sequencing and transcriptome assembly

For deep sequencing of *C. plantagineum* transcriptomes, four cDNA libraries were constructed, one each from mRNA of leaf tissue of fully hydrated plants (control), dehydrated for 48 h (80% RWC; early dehydrated), dehydrated for 15 days (5% RWC; desiccated) and re-hydrated for 24 h (rehydrated). These four samples represent critical physiological changes during the dehydration/rehydration cycle of *C. plantagineum* as illustrated in Figure 9.1. After preparation of the cDNA libraries and

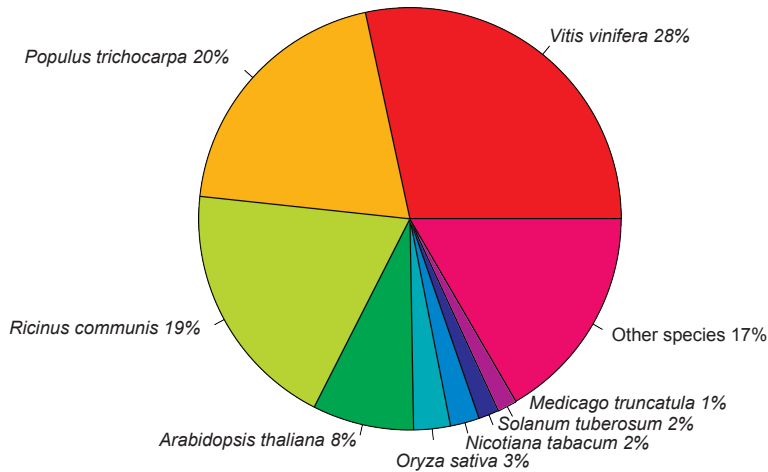




**Figure 9.1** – Four stages in the dehydration/rehydration cycle of *Craterostigma plantagineum* from which expressed sequence tag (EST) libraries were constructed. (a) Fully hydrated plant (control). (b) Dehydrated for 48 h (80% relative water content, RWC). (c) Desiccated (5% RWC). (d) Rehydrated for 24 h.

their pyro-sequencing, a total of 182 Mb of transcripts were obtained that were assembled into 29 400 contigs, spanning a total of 12.8 Mb of sequence. More than 15 000 UniProt identities were then found by performing a BLAST analysis of the assembled sequences (File S1 in Supporting Information). About a third of the contigs did not map to UniProt identities, and therefore represent a source for gene discovery. These sequences were not included in the analysis described below (see Experimental Procedures for additional details).

The BLAST analysis revealed that a large portion of *C. plantagineum* sequences are similar to sequences found in the recently published grape (*Vitis vinifera*) genome.<sup>14</sup> Sixty-seven per cent of the *C. plantagineum* transcripts are more similar to *V. vinifera*, castor bean (*Ricinus communis*) and poplar (*Populus trichocarpa*) than to any other species. The total distribution of closely related sequences from different plants and other organisms found in the *C. plantagineum* transcriptome is shown in Figure 9.2. Although we have only included the top matches, these results suggest that a large proportion of *C. plantagineum* transcript contigs do not have close relatives in model plants such as *Arabidopsis thaliana*. Therefore, studying the *C. plantagineum* transcriptome may provide new insights into the mechanisms of desiccation in more closely related plants such as grape.



**Figure 9.2** – Species distribution of the UniProt identities from the BLAST analysis of *Craterostigma plantagineum* transcriptomes with expected  $E$ -value  $< 10^{-4}$ . More than 60% of the identified transcripts have highest similarity with *Vitis*, *Ricinus* or *Populus*

### Transcript clustering by expression signatures

In order to analyze which cellular processes are critical during the different hydration stages, transcripts were grouped by their expression signatures across the four samples. This was done by partition clustering the 500 most variable transcripts according to their profiles, which identified six clusters representing gene sets of specific expression signatures (Figure 9.3, Table S1). Cluster I contains the 159 most abundant transcripts in the control sample only. Cluster II includes the 112 most abundant transcripts in the early dehydrated sample exclusively. Cluster III contains the 88 transcripts most abundant in the desiccated sample. Cluster IV has 88 transcripts most abundant in the rehydrated sample. The smaller clusters V and VI contain transcripts more abundant in both control and rehydrated samples and the transcripts with higher expression levels in both early dehydrated and desiccated samples, respectively. The most abundant transcripts for each cluster are listed in Table 9.1. Representative UniProt IDs are described in the following sections.

Cluster	UniProt ID	Control	Dehyd.	Desicc.	Rehyd.	Description
I	Q7X999	2749.61	132.15	83.59	589.47	Ribulose biphosphate carboxylase/activase
	Q9XGN4	535.13	62.21	30.39	46.99	Galactinol synthase OS= <i>Ajuga reptans</i>
	Q5ZF64	238.03	71.92	30.39	40.07	Cysteine protease 1 OS= <i>Plantago major</i>
	B2BGT2	227.05	26.12	30.39	84.14	Putative ribulose-1,5-bisphosphate carboxylase
	A9PJH1	226.51	3.86	3.84	20.13	Carbonic anhydrase
	B9SUE9	152.02	21.37	25.28	43.45	Hydrolase, hydrolyzing O-glycosyl compounds
	Q8L5E1	139.28	3.86	6.21	5.85	Acid phosphatase OS= <i>Lupinus luteus</i>
	B9T0C1	139.28	50.44	5.35	36.81	Chlorophyll A/B binding protein
	A7PND1	130.42	23.03	22.92	24.82	Chromosome chr1 scaffold_22 OS= <i>Vitis vinifera</i>

(Continued on next page)

Table 9.1 – (continued)

Cluster	UniProt ID	Control	Dehyd.	Desicc.	Rehyd.	Description
II	Q5ZF87	109.97	33.92	10.53	22.77	Putative uncharacterized protein OS= <i>P. major</i>
	B9S8X4	105.57	18.38	3.84	3.63	Ribonuclease t2 OS= <i>Ricinus communis</i>
	Q43042	100.07	14.77	6.21	17.36	Nitrate reductase apoenzyme OS= <i>Petunia hybrida</i>
	B9T2G5	97.88	8.7	3.84	18.16	Fructose-1,6-bisphosphatase
	A7R0A2	91.85	17.47	6.21	34.15	Chromosome chr14 scaffold301 OS= <i>V. vinifera</i>
	O64445	85.31	7.64	5.35	25.47	Light harvesting chlorophyll a/b-binding protein
	A9PEX2	9.15	382.39	92.64	52.2	Predicted protein OS= <i>Populus trichocarpa</i>
	Q9ZPC6	26.85	375.42	74.46	41.4	Sucrose synthase OS= <i>C. plantagineum</i>
	Q9FUG0	16.62	314.21	97.49	41.4	Lea1P OS= <i>Daucus carota</i>
	A7PR86	5.54	280.11	33.41	25.47	Chromosome chr14 scaffold_26 <i>V. vinifera</i>
	B1Q190	78.81	239.3	34.18	19.03	Pathogenesis-related protein
	Q2QKE8	7	237.35	66.71	38.76	Late embryogenesis abundant protein
	O23764	6.29	194.23	53.23	29.32	CDet11-24 protein OS= <i>C. plantagineum</i>
	O16527	8.44	172.53	37.96	15.73	Protein K08H10.1, Lea
	Q38JC5	5.54	166.6	33.41	50.69	Temperature-induced lipocalin
	B9RMC7	3.72	156.03	40.18	16.61	Putative uncharacterized protein OS= <i>R. coicinus</i>
	Q75VR0	7.72	151.4	27.64	11.11	Two pore calcium channel protein
	P22238	3.72	150.08	31.88	8.79	Desiccation-related PCC27-04 OS= <i>C. plantagineum</i>
	B9GN09	9.15	149.42	13.04	13.36	Predicted protein OS= <i>P. trichocarpa</i>
	Q84P54	9.15	136.15	25.28	19.63	Gamma aminobutyrate transaminase isoform1
	A7NV27	9.92	135.48	36.47	28.01	Chromosome chr18 scaffold_1 <i>V. vinifera</i>
	Q645Q9	9.15	239.3	723.46	103.57	Late embryogenesis-like protein
	Q93XQ9	25.6	46.98	519.17	88.82	Putative cysteine protease OS= <i>Ipomoea batatas</i>
	Q8H6D8	6.29	211.91	477.78	84.92	LEA1-like protein OS= <i>C. annuum</i>
	A6N8F8	26.85	22.19	335.28	60.65	Cysteine proteinase
B9RV72	14.49	5.57	294.43	49.2	Delta 9 desaturase, putative	
Q9LT78	13.82	25.37	284.5	59.87	Cysteine proteinase OS= <i>Arabidopsis thaliana</i>	
P22241	6.29	89.8	265.25	34.82	Desiccation-related PCC27-45 <i>C. plantagineum</i>	
Q5ZF62	9.92	30.28	260.59	48.46	Cysteine protease 3 (Fragment)	
B9RV37	22.68	32.47	192.25	46.27	Lactoylglutathione lyase	
B9HWU0	28.12	29.56	190.23	19.03	Predicted protein OS= <i>P. trichocarpa</i>	
A7QFJ2	11.25	19.27	180.78	79.43	Chromosome chr8 scaffold_88, <i>V. vinifera</i>	
A7Q8Q2	53.49	19.27	141.42	32.08	Chromosome chr5 scaffold_64, <i>V. vinifera</i>	
A7QA45	17.81	8.7	135.28	28.66	Chromosome scaffold_69, <i>V. vinifera</i>	
B9I1Q4	9.15	11.86	130.5	19.03	Predicted protein OS= <i>P. trichocarpa</i>	
A5AM73	3.72	3.86	130.5	14.92	Putative uncharacterized protein OS= <i>V. vinifera</i>	
IV	A6N8C4	36.19	24.61	26.14	407.42	Pathogen-related protein STH-2
	A7P3T0	43.37	26.83	64.59	255.6	Chromosome chr1 scaffold_5, <i>V. vinifera</i>
	Q5QIS5	63.49	27.51	91.95	254.86	Ascorbate peroxidase OS= <i>Rehmannia glutinosa</i>
	B5KVN9	52.52	21.37	4.57	233.39	Pathogenesis related protein PR10
	A5B5A4	59.43	26.12	50.36	232.65	Putative uncharacterized protein, <i>V. vinifera</i>
	B9IG30	26.85	10.85	34.95	205.85	Predicted protein OS= <i>P. trichocarpa</i>
	A7P406	50.11	31	19.51	205.1	Chromosome chr1 scaffold_5, <i>V. vinifera</i>
	B9RY80	23.28	8.7	4.57	199.87	Putative uncharacterized protein OS= <i>R. communis</i>
	Q9SSZ7	17.22	4.66	5.35	194.63	Peroxidase 3 OS= <i>Scutellaria baicalensis</i>
	Q2YHM9	35.65	7.64	12.23	169.09	Caffeoyl-CoA O-methyltransferase, <i>P. major</i>
	A7PNX9	18.99	7.64	27.64	133.49	Chromosome chr8 scaffold_23, <i>V. vinifera</i>
	A6YGE4	15.96	14.77	35.71	127.39	Hypersensitive-induced response protein
	B9RY25	13.82	12.83	66	122.8	Catalytic, putative OS= <i>R. communis</i>
	A7P3S7	29.37	6.58	41.66	102.03	Chromosome chr1 scaffold_5, <i>V. vinifera</i>
	B1Q468	17.81	5.57	8.79	101.25	Flavonoid glucosyltransferase, <i>Antirrhinum majus</i>
	Q9XQB5	831.91	118.13	27.64	799.01	Ribulose biphosphate carboxylase small chain
	Q3EBJ5	275.8	23.03	33.41	121.27	Uncharacterized protein At2g39730.3
	B9MVC6	251.19	13.77	3.84	120.5	Predicted protein OS= <i>P. trichocarpa</i>
	Q9ZRI4	230.35	13.77	7.97	170.59	Ribulose biphosphate carboxylase small chain
	O23772	159.22	24.61	9.63	88.82	Major intrinsic protein PIPC OS= <i>C. plantagineum</i>
	A7QBQ2	154.24	16.68	9.63	56.78	Lipoxygenase OS= <i>V. vinifera</i>
	Q40271	117.7	11.86	15.56	63.76	Inositol-3-phosphate synthase
	Q8L5T3	92.94	11.86	9.63	45.56	Rubisco activase OS= <i>Chenopodium quinoa</i>
	Q9SA52	89.66	15.81	7.97	41.4	Uncharacterized protein At1g09340
	B9SMK7	88.03	13.77	6.21	33.47	Serine hydroxymethyltransferase
B9IF13	86.39	6.58	5.35	35.49	Predicted protein OS= <i>P. trichocarpa</i>	
Q2LAH0	84.76	7.64	3.84	38.76	Chloroplast photosystem II 22 kDa component	
A5A7P7	80.43	5.57	11.36	58.33	Glutamine synthetase OS= <i>Avicennia marina</i>	
B9T1G0	78.81	9.76	9.63	33.47	Magnesium-chelatase subunit H	
B9HMK4	73.98	4.66	10.53	49.2	Predicted protein (Fragment) OS= <i>P. trichocarpa</i>	
VI	P22242	12.57	1510.6	1296.7	126.62	Desiccation-related PCC13-62 OS= <i>C. plantagineum</i>
	P23283	17.81	1135.2	450.86	176.62	Desiccation-related PCC3-06 OS= <i>C. plantagineum</i>
	Q84RM1	17.81	704.54	367.35	55.24	Small blue copper protein Bcp1
	Q42671	7	377.32	196.3	34.15	Glyceraldehyde-3-phosphate dehydrogenase, <i>C. plantagineum</i>
	Q39284	3.72	142.12	75.17	15.73	Aldose reductase-like protein OS= <i>Bromus inermis</i>
	P22239	6.29	134.82	95.42	14.13	Desiccation-related PCC6-19 OS= <i>C. plantagineum</i>
	Q9FLN9	3.72	99.97	46.03	18.16	Embryo-specific protein 1; Ca2+-binding EF-hand
	B9IDA6	6.29	87.75	43.85	20.7	Predicted protein OS= <i>P. trichocarpa</i>
	Q9SQ57	3.72	82.96	45.3	16.61	Caleosin OS= <i>Sesamum indicum</i>
	Q42908	4.25	72.61	33.41	9.54	bisPO4glycerate-independent PO4glycerate mutase
Q42620	5.54	68.45	53.95	11.81	CTP:phosphocholine cytidylyltransferase	

(Continued on next page)

Table 9.1 – (continued)

Cluster	UniProt ID	Control	Dehyd.	Desicc.	Rehyd.	Description
	B9MT96	3.72	65.68	30.39	14.13	Predicted protein OS= <i>P. trichocarpa</i>
	A7QFG4	3.72	63.6	38.7	9.54	Chromosome chr8 scaffold_88, <i>V. vinifera</i>
	Q9SBR9	5.54	60.14	73.76	8.09	Basic blue protein OS= <i>Medicago varia</i>
	Q3YMR1	3.72	41.49	46.03	6.57	Class 4 LEA protein OS= <i>C. plantagineum</i>

Table 9.1 – The most differentially expressed transcripts in leaves of *Craterostigma plantagineum* during the dehydration/rehydration cycle. The table shows the total read count for each gene after normalization across the four samples: (a) control, (b) dehydrated, (c) desiccated, (d) rehydrated.

### Gene ontology groups in dehydrated and rehydrated tissues

GO enrichment of each of the six expression signature clusters of genes identified the biological processes (BP) and metabolic functions (MF) that characterize each cluster (Table 9.2 and Table S2). Phenotypic profiles were obtained via the GO-enriched categories that were significantly represented in each cluster. Such profiles are indicative of those physiological functions that are regulated at the transcript level during the dehydration/rehydration cycle. The control sample contains a high representation of transcripts involved in cell wall construction and organization, as well as in normal growth and homeostasis. Carbohydrate metabolism is enriched in this cluster, together with oxidoreductases and enzymes related to

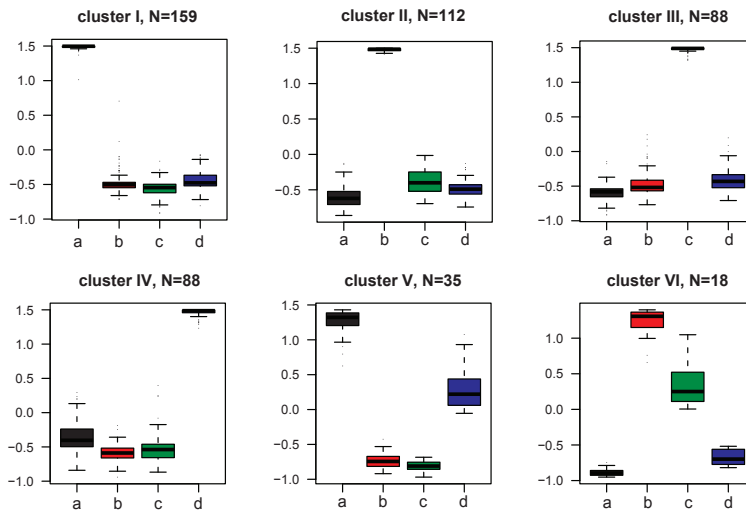


Figure 9.3 – Six clusters representing expression signatures in the four stages of *Craterostigma plantagineum*. *n* is the number of UniProt transcripts found in each cluster. The y-axis represents normalized expression values. The expression of each gene was scaled to a mean of 0 and standard deviation of 1 across samples. Samples: (a) control, (b) dehydrated, (c) desiccated, (d) rehydrated.

oxylipin metabolism that may participate in the lipoxygenase (LOX) pathway. The profile of abundant transcripts in the early dehydrated sample indicates predominant expression of proteins involved in ABA-mediated responses to water stress and general abiotic and biotic stresses including ion homeostasis-related proteins. These results coincide with previous studies assigning ABA-related processes a central role in early desiccation events.<sup>12</sup> In addition, cluster II showed a significant representation of transcripts related to thiamine metabolism. Such processes have been proposed to protect plants from oxidative stress<sup>15</sup> as the accumulation of ROS is a general stress response that can be triggered by osmotic and ionic imbalance in water-stressed cells. Remarkably, desiccated samples were characterized by a predominance of transcripts involved in tryptophan metabolism, as well as in processes related to the metabolism of indole derivatives seen in cluster III. These biological functions may be involved in the salvage of precursors for the metabolism of proteins and nucleic acids in the desiccated, quiescent stage.

The profile of transcripts enriched in rehydrated samples (cluster IV) mainly represents proteins protective against oxidative stress and enzymes involved in the metabolism of vitamin K-related compounds. Such processes could enable a gradual recovery of regular metabolic functions as water re-enters cells. Cluster V includes transcripts which are similarly enriched in both control and rehydrated tissue samples. The functions of the corresponding genes are required for photorespiration and photosynthesis, as well as intracellular transport and assembly of protein complexes with the participation of Rubisco, GTPases and carbon utilization enzymes. Lastly, cluster VI includes transcripts differentially upregulated in early dehydrated and desiccated samples and associated with a general stress response.

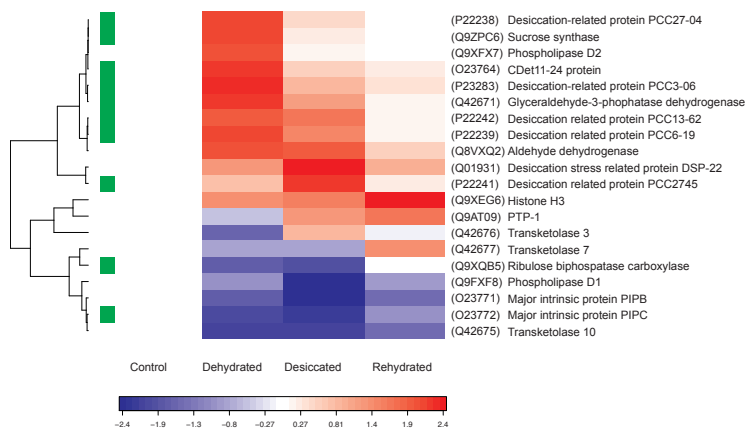
### Confirmation of gene expression profiles

Previous studies quantified the expression of a number of *C. plantagineum* genes at different stages of hydration and dehydration.<sup>16–18</sup> Twenty such genes with previously reported differential expression were selected *a priori* and assessed in our data set. This provides a rough estimate of the accuracy of our assembly of transcript reads, the UniProt annotations from BLAST, and the expression profiles. To evaluate whether the 20 genes were significantly differentially expressed in our data, we performed a hypergeometric test on the top 500 genes showing an overlap for 10 of the 20 genes with the 500 which resulted in a  $P$ -value of  $6.1 \times 10^{-12}$ , and a ‘gene set enrichment analysis’ test on all 20 genes resulting in a  $P$ -value of  $2.7 \times 10^{-9}$ . Additionally, a hierarchical cluster analysis of normalized expression values for the 20 genes was performed (Figure 9.4). This analysis showed correspondence between our data and the previous analysis of this set of transcripts. Both our data and those of Bernacchia et al. (1996)<sup>16</sup> showed that abundant transcripts in control and rehydrated tissues correspond to Rubisco small subunit Q9XQB5 that gradually decreases as tissues undergo dehydration. Genes encoding LEA proteins, including *pcC3-06* (P23283), *pcC27-04* (P22238) and *CDeT11-24* (O23764), have comparable patterns of expression, being more abundant upon early dehydration.<sup>19</sup> A similar profile includes *ppC13-62* (P22242) and *pcC6-19* (P22239) with higher expression in

Cluster	BP	MF
I	Plant-type cell wall organization	Xyloglucan:xyloglucosyl transferase activity
	Cellular glucan metabolic process	Oxidoreductase activity
	Cellular polysaccharide metabolic process	Galactinol-raffinose galactosyltransferase
	Nitrate metabolic process	Lipoxygenase activity
	Oxylipin biosynthetic process	Endoribonuclease activity
		Transferase activity, transferring hexosyl
		Hydrolase activity, hydrolyzing O-glycosyl
		Carbonate dehydratase activity
		Galactosyltransferase activity
		Lactoylglutathione lyase activity
II	Thiamin biosynthetic process	Ferric iron binding
	Response to water	Phosphotransferase activity
	Response to abiotic stimulus	
	Iron ion transport	
	Response to abscisic acid stimulus	
III	Cellular iron ion homeostasis	
	Response to desiccation	Cysteine-type endopeptidase activity
	Tryptophan metabolic process	Tryptophan synthase activity
IV	Indolalkylamine metabolic process	Oxidoreductase activity, acting on paired donors
		Cysteine-type peptidase activity
	Menaquinone metabolic process	Naphthoate synthase activity
	Fat-soluble vitamin metabolic process	Calcium ion binding
	Vitamin K biosynthetic process	Oxo-acid-lyase activity
	Quinone cofactor metabolic process	Peroxidase activity
	Response to oxidative stress	Oxidoreductase activity, acting on peroxide
	Methionine biosynthetic process	Calcium-dependent phospholipid binding
V		S-methyltransferase activity
	Photorespiration	Antioxidant activity
	Photosynthesis	Heme binding
	Carbon utilization by fixation of CO <sub>2</sub>	Ribulose-bisphosphate carboxylase activity
	Protein polymerization	
	Cytoskeleton-depend. intracell transport	
VI	Response to stress	Copper ion binding

**Table 9.2** – Gene ontology (GO)-enriched categories of six expression signature clusters for UniProt IDs under biological processes (BP) and metabolic functions (MF). GO terms significantly over-represented in control sample (I), dehydrated (II), desiccated (III) and rehydrated (IV). Cluster V contains more abundant transcripts in control and rehydrated samples, cluster VI abundant transcripts in dehydrated and desiccated stages. GO terms under BP are independent of MF categories.

tissues undergoing dehydration.<sup>20</sup> The LEA-like gene *pcC27-45* (P22241), belongs to a related cluster with higher expression in desiccated tissues.<sup>21</sup> Our data also agree with the finding of Kirch et al. (2001)<sup>22</sup> that mRNA of the aldehyde dehydrogenase *CpALDH* (Q8VXQ2) accumulated to high levels in dehydrated, ABA-treated tissues. In addition, our profiling confirmed expression patterns reported for two *C. plantagineum* phospholipase genes in leaves.<sup>23</sup> Expression of *CpPLD1* (Q9XFX8) was high in hydrated conditions but decreased upon dehydration, whereas *CpPLD2* (Q9XFX7) was more abundant in early dehydrated tissues. This agrees with



**Figure 9.4** – Expression patterns of 20 previously characterized *Craterostigma plantagineum* transcripts. Colors indicate expression values scaled to standard deviations and centered at the control intensity level (Z-score). Red indicates increased expression and blue decreased, relative to the control conditions. The green bars indicate transcripts in the top 500 differentially expressed ones in our data.

the proposal that *CpPLD2* is involved in early dehydration signaling. The expression of members of the family of transketolase (*tkt*) genes in *C. plantagineum* was also analyzed. *Tkt3* (Q42676) was reported to be constitutively expressed at low levels, whereas *tkt7* (Q42677), like *tkt10* (Q42675), was more abundant in rehydrating tissues.<sup>24</sup> Our transcriptional profiling coincides for *tkt7*, while *tkt10* was more abundant in control tissues and *tkt3* was also expressed in desiccated leaves.

### Metabolic pathways in the dehydration-rehydration cycle

GO enrichment identified metabolic pathways which may be important for desiccation tolerance. To further delineate such metabolic pathways, the Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.genome.jp/kegg><sup>25</sup>) and the MetaCyc Encyclopedia of Metabolic Pathways (<http://metacyc.org><sup>26</sup>) databases were mapped to the annotations in our transcript data. As the annotation coverage of UniProt transcripts in our assembly was poor in these databases, putative Arabidopsis ortholog transcripts were used. This revealed significantly enriched pathways (Table S3): photosynthesis, phenylpropanoid biosynthesis, photosynthetic carbon fixation, one carbon pool by folate generating C1 units for photorespiration and photosynthesis,<sup>27</sup> phytohormone biosynthesis, methane metabolism, ubiquinone and other terpenoid-quinone biosynthesis, geraniol degradation and the superpathway of geranylgeranyldiphosphate biosynthesis II (via methylerythritol phosphate, MEP). Expression patterns within these metabolic pathways were then analyzed by hierarchical clustering, and four of those pathways are shown in Figure 9.5. Heatmaps of the pathways are shown with the expression

profiles of identified transcripts. Differential expression of transcripts encoding proteins involved in the pathways depicted is a hallmark of each of the four stages of hydration and dehydration.

Discrete steps of individual pathways were visualized by mapping the components of metabolic pathways to Arabidopsis orthologs in the KEGG database. Differential expression is indicated by a color scale that reflects the coefficient of variation. The resulting colored pathway figures indicate putative nodes of transcriptional regulation across the profiles of the four samples. Figure 9.6 illustrates the biosynthesis of plant hormones and shows that late steps in ABA, gibberellin (GA), ethylene and jasmonic acid (JA) synthesis appear to be transcriptionally regulated. Other examples of mapped metabolic profiles can be visualized online by loading File S2 ([http://www.genome.jp/kegg/tool/color\\_pathway.html](http://www.genome.jp/kegg/tool/color_pathway.html)).

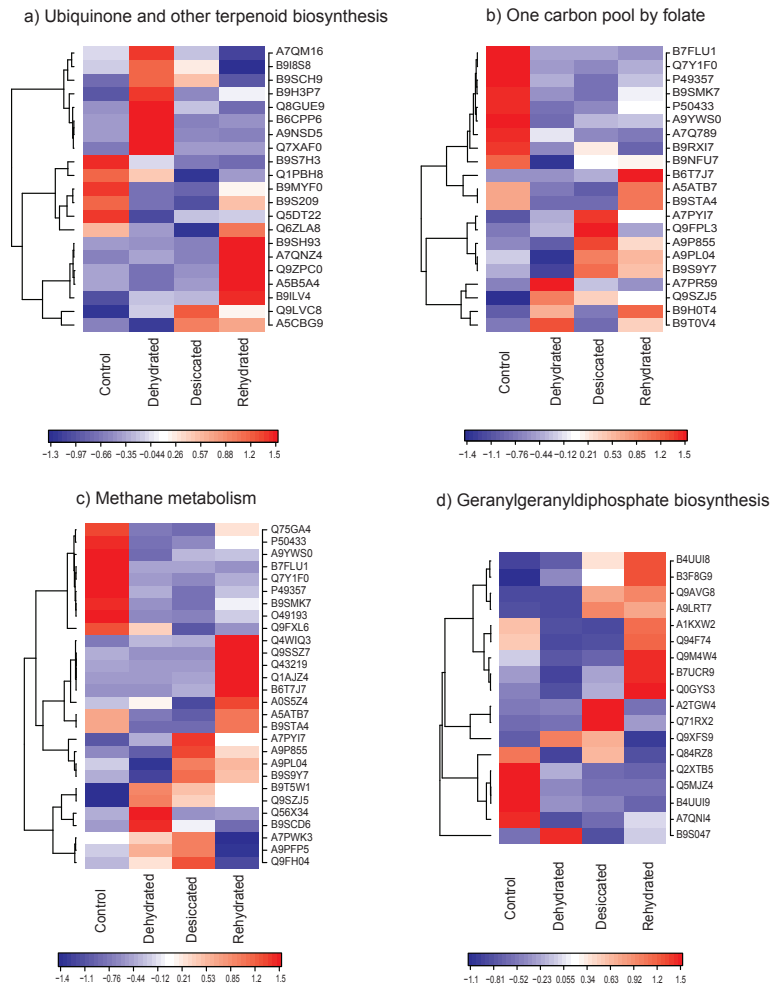
### Expression profiles in *C. plantagineum* and desiccationtolerant seeds

Complementary DNA library subtraction has been used to compare gene expression in developing barley embryos and *C. plantagineum* dehydrating tissues.<sup>20</sup> This study showed that most cDNAs encoded LEA proteins, other ABA-induced proteins and osmolyte-producing enzymes. Concurrent with our transcriptomics analysis, similar pathways provide protection upon desiccation stress in the vegetative tissues of *C. plantagineum*. All LEAs and general dehydration-related UniProt IDs were filtered from our data and their Arabidopsis orthologs were BLAST searched. Comparison of their expression patterns showed that putative orthologs of predominantly seed localized Arabidopsis LEAs (*At3g15670*, *At3g53040*, *At2g36640*, *At4g21020*, *At5g44310*) are found in *C. plantagineum* vegetative tissues during dehydration stress. Table 9.3 shows the expression levels of Arabidopsis putative orthologs of *C. plantagineum* LEAs and other desiccation-related proteins. Some of these Arabidopsis genes are up-regulated upon osmotic or salt stress, but not during dehydration.<sup>28</sup> This comparison indicates that desiccation tolerance in resurrection plants involves many genes implicated in seed desiccation tolerance. This is further corroborated by comparison of our data with those of Boudet et al. (2006)<sup>29</sup> and Buitink et al. (2006)<sup>30</sup> who describe 64 genes associated with desiccation tolerance in Medicago seeds. Fifty of these 64 Medicago transcripts had significant UniProt orthologs among the Craterostigma transcripts analyzed here. Four LEA proteins (*TC85220*, *TC87163*, *TC76866*, *TC78559*) described by Buitink et al. (2006)<sup>30</sup> are among the 500 most highly expressed genes among the early dehydrated Craterostigma genes (Table S4, hypergeometric *P*-value = 0.02).

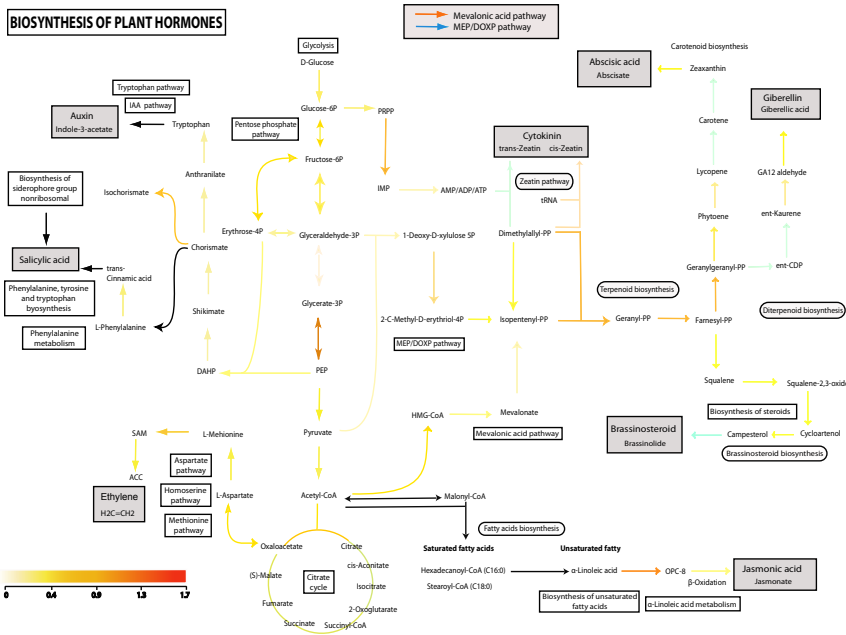
### Expression profiles in dehydrating *C. plantagineum*, Arabidopsis and Vitis

Gene expression in response to dehydration was compared between *C. plantagineum* and air-dried Arabidopsis vegetative tissues which are desiccation intolerant.<sup>31</sup> This revealed that LEA orthologs abundant in dehydrating *C. plantagineum* had increased expression in response to dehydration in Arabidopsis. We also compared our data with transcriptional responses to dehydration in Vitis.<sup>32</sup> Interestingly, the profiles of genes encoding Rubisco activase, cystatin, galactinol synthase





**Figure 9.5** – Representative metabolic pathways in *Craterostigma plantagineum* transcript profiles. Heat maps of patterns of expression corresponding to four of the most enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways. The heat scale shows normalized and Z-scaled expression (mean centered, with standard deviations as unit). Red indicates increased expression and blue decreased, relative to the four conditions



**Figure 9.6** – Metabolic map of phytohormone biosynthetic pathways in *Craterostigma plantagineum* putative Arabidopsis orthologs. The color key indicates the coefficient of variation as a measure of differential expression of the transcripts across the four hydration/dehydration stages.

or expansins characterize them as dehydration responsive genes in *Vitis*. In *C. plantagineum*, Rubisco activase and galactinol synthase transcripts are abundant in control plants, suggesting a constitutive priming to respond to dehydration (Table 9.1). In terms of phytohormone-related responses, the studies in *Vitis* established that ABA-related transcripts showed differential expression, followed by transcripts related to GA metabolism. Our profiles also show regulation in the synthesis and responses to both hormones in accordance with the GO enrichment and metabolic pathways (see above).

UniProt ID	Control	Dehyd.	Desicc.	Rehyd.	Description	Putative Ortholog	Dev. Stage	Abiotic stress
P22242	12.57	1510.61	1296.75	126.62	PCC13-62	At1g47980	fl	os, ox
P23283	17.81	1135.23	450.86	176.62	PCC3-06	At4g21020	fl, imb	os
Q01931	9.92	1069.88	2082.01	826.54	DSP-22	At4g14690	fl	co, uv
Q9FUG0	16.62	314.21	97.49	41.4	LEA 1P	At1g52690	fl	os, co, sa
Q645Q9	9.15	239.3	723.46	103.57	LEA-like	At2g46140	fl, ro	all
Q2QKE8	7	237.35	66.71	38.76	LEA	At3g15670	fl, imb	os, co, sa
O23764	6.29	194.23	53.23	29.32	CDet11-24	At5g52300	p	os, co, sa
O16527	8.44	172.53	37.96	15.73	LEA	At1g72100	p, imb	os
Q9FLN9	3.72	99.97	46.03	18.16	EP	At5g55240	fl	os, co, sa
P22241	6.29	89.8	265.25	34.82	PCC27-45	At1g01470	am	os, co, sa
B5TV63	15.96	74	32.64	17.36	SRP	At2g17840	imb	os, co, sa
B9RBC1	4.83	36.04	8.79	9.54	LEA	At5g44310	fl, imb	os
Q9LF88	4.83	22.19	11.36	8.09	LEA-like	At3g53040	p, imb	co
Q9LLQ6	3.72	20.63	6.21	5.85	SMP PM34	At1g54870	fl, imb	os, sa
P20075	3.72	9.76	7.97	5.16	EP DC-8	At2g36640	fl	os, co, sa
B1PB87	3.72	7.64	3.84	4.52	LEA1	At5g06760	fl, imb	os, co, sa
Q9XES7	3.72	4.66	6.21	3.63	SMP PM27	At1g22600	fl, imb	

**Table 9.3** – Comparison of expression profiles of dehydration-related proteins in *Craterostigma plantagineum* putative orthologs from Arabidopsis. Expression measures are the normalized read count values across the four *C. plantagineum* samples: (a) control, (b) dehydrated, (c) desiccated, (d) rehydrated. SRP, senescence-related protein; SMP, seed maturation protein; EP, embryonic protein; Dhn, dehydrin. Expression patterns from microarray experiments in different developmental stages, during germination and in response to abiotic stresses compiled by AtGenExpress ([jsp.weigelworld.org/expviz/expviz.jsp](http://jsp.weigelworld.org/expviz/expviz.jsp)): floral organs (fl), pollen (p), roots (ro), apical meristem (am), osmotic stress (os), oxidative stress (ox), cold (co), salt (sa), all stresses (all), imbibition (imb), uv (uv irradiation).

## Discussion

Expressed sequence tag (EST) analysis is a rapid and efficient method for identifying genes involved in specific metabolic processes, especially in organisms lacking a genome sequence. Small-scale cDNA sequencing projects have been performed in resurrection plants such as *Tortula ruralis*,<sup>33,34</sup> *Selaginella lepidophylla*,<sup>35</sup> *Sporobolus stapfianus*<sup>36</sup> and *C. plantagineum*.<sup>17</sup> Although resurrection plants may be a source of genetic determinants for drought tolerance, gene discovery efforts in these plants have been limited thus far. Our largescale EST sequencing analysis identified some 15 000 putative transcripts which mapped to UniProt. The expression patterns of the transcripts were characterized in the leaves of *C. plantagineum* at four physiological stages during a cycle of dehydration and rehydration.

### Metabolic regulation in desiccation-tolerant *C. plantagineum*

Abundant transcripts in the unstressed control sample encode proteins involved in photosynthesis and polysaccharide metabolism. The most abundant transcripts in this cluster encode Rubisco activase, carbonic anhydrases, galactinol synthases and fructose biphosphatases. Galactinol synthases catalyze the first committed step in the synthesis of raffinose oligosaccharides (RFOs). Accumulation of RFOs has been

associated with cold, heat or dehydration stress<sup>37</sup> and with acquisition of desiccation tolerance in orthodox seeds.<sup>38,39</sup> For example, RFOs accumulate during developmental desiccation and later at the onset of imbibition in tomato seeds.<sup>40</sup>

Other transcripts noticeably present in the unstressed control encode acid phosphatases which are required to maintain inorganic phosphate levels.<sup>41</sup> Phosphorus availability and carbohydrate metabolism are connected, as inorganic phosphorus (Pi) starvation causes down-regulation of genes involved in photosynthesis and nitrogen and carbon metabolism.<sup>42</sup> In accordance with this, in our GO enrichment analysis there is significant representation in the control sample of transcripts involved with nitrate metabolism. These results indicate that N–Pi balance is under transcriptional control in *C. plantagineum*.

The unstressed sample also appears to be enriched in transcripts encoding proteins involved in ion transport, such as membrane-associated carriers, together with proteins involved in the maintenance of cell wall plasticity and membrane integrity such as xyloglucan endotransglucosylases and expansins.<sup>43</sup> The highly abundant *C. plantagineum* expansin A2 described earlier is thought to play a role in cell wall flexibility to withstand water depletion.<sup>44</sup> Such flexibility may be required to avoid desiccation-related damage due to mechanical strain in cells undergoing shrinkage. In *Craterostigma* species, responses such as the redistribution of Ca<sup>2+</sup> and xyloglucan modification with the intervention of expansins has been observed.<sup>44,45</sup> Morphological studies have reported the appearance of cell wall folding in desiccated *Craterostigma* plants. Xyloglucan and pectin-modifying enzymes could account for this capacity to remodel cell wall composition and for maintaining cell wall flexibility required for water loss and up-take.

### **The dehydration transcriptome – ABA-mediated responses and thiamin metabolism**

Cluster II includes transcripts pre-eminently expressed during early dehydration of *C. plantagineum*. The physiological condition at the early dehydrated state is comparable in terms of water content with dehydration treatments studied in non-desiccation tolerant plants such as *Arabidopsis*. The early dehydrated group of *Craterostigma* transcripts contains transcripts related to carbohydrate metabolism such as sucrose synthase, as well as dehydrins and other groups of LEA proteins expressed in response to water stress. Thiamin biosynthetic processes were highlighted in our transcript profile, in agreement with recent evidence linking thiamin-related metabolism and oxidative stress responses in plants.<sup>15</sup> For example, thiamin treatment leads to the induction of systemic acquired resistance (SAR). This type of priming is dependent on ROS and the NPR1 protein that regulates responses to pathogens.<sup>46–48</sup> It is possible that thiamine metabolism plays protective roles against several stressors. For example, thiamine pyrophosphate (TPP) is a co-factor for enzymes that regulate carbohydrate and amino-acid metabolism.<sup>49</sup> In addition, thiamine biosynthesis and expression of the thiazole-synthesizing gene *thi1* increase in *Arabidopsis* in response to flooding, salinity and reduced sugars. In maize (*Zea mays*) seedlings, abiotic stress prompts accumulation of free thiamine, possibly for later use in thiamine diphosphate (TDP)-dependent processes.<sup>50,51</sup> TDP-

dependent processes enable the accumulation of soluble sugars for energy generation required for synthesizing stress-related proteins or osmoprotectants.<sup>52</sup> In several plants, including *Arabidopsis*, the TDP-dependent enzyme pyruvate decarboxylase is induced by anoxia, low temperature, osmotic stress, ABA, oxidative stress and wounding.<sup>53,54</sup> Accordingly, our profiling indicates that the metabolic route for producing pyruvate is differentially regulated in *C. plantagineum* (Figure 9.6).

Another example of TDP-dependent enzymes regulated at the transcript level in *C. plantagineum* are the transketolases.<sup>55</sup> These are key enzymes of the pentose phosphate cycle in all organisms and in the reductive part of the Calvin cycle in plants. Expression profiles of *C. plantagineum* transketolases have previously been associated with unusual carbohydrate metabolism.<sup>24</sup> Transketolases are thought to participate in the synthesis of octulose, a rare sugar that accumulates in *C. plantagineum*.<sup>56</sup> Our results support the previous evidence that the production of thiamine and related compounds enables *C. plantagineum* to tailor specific aspects of the response of carbohydrate metabolism to desiccation.

A further type of stress-responsive transcript that accumulates during dehydration in *C. plantagineum* encodes temperature-induced lipocalins. These membrane-associated proteins have low-temperature response elements (LTREs), dehydration response elements (DREs) and heat shock elements (HSEs) in their promoter regions.<sup>57</sup> Temperature stress can cause membrane damage, and membrane-anchored lipocalins appear to be expressed under conditions that affect membrane integrity. In animals, lipocalins bind steroid hormones and cholesterol. Sterol binding may enhance tolerance to extreme temperatures by allowing increased membrane fluidity and maintaining phospholipid arrangements.<sup>58</sup> Lipocalins may also function in the biosynthesis of xanthophylls required to protect photosystem II (PSII) against photo-oxidative stress.<sup>59</sup> In line with other temperature and dehydration responsive genes, we found that transcripts of the COR413 thylakoid membrane targeted protein accumulate under dehydration. Proteins of this family are thought to play roles in membrane stabilization or in stress signaling via association with membrane receptors.<sup>60</sup>

Other transcripts in the dehydration sample encode aquaporins, tonoplast intrinsic proteins, cation transporters and rare cold inducible 2A (RCI2A<sup>61</sup>). Genes with regulatory functions were also identified such as NAC transcription factors that are induced by drought and salinity<sup>62</sup> and a homolog of abscisic acid insensitive 1 (ABI1) encoding a phosphatase PP2C regulating early ABA responses.<sup>63</sup> In line with this, ABA signaling is important for desiccation tolerance in early stages of dehydration and in calli of *C. plantagineum*.<sup>8</sup> In general, accumulation of ABA-regulated transcripts is also an early response to dehydration in *C. plantagineum*.

### The desiccated transcriptome

Transcripts of desiccated plants are dominated by those encoding LEA proteins, DNA-binding proteins, cysteine proteases and proteins of DNA and amino-acid metabolism. The accumulation of transcripts for several types of cysteine proteases is intriguing. Such proteases are found in tissues undergoing oxidative stress-mediated programmed cell death (PCD<sup>64</sup>). It is therefore possible that water loss

triggers their expression to initiate a cellular recycling program. In addition, cysteine proteinases are expressed during seed development and germination to degrade storage proteins.<sup>65</sup> These transcripts may also be proactively accumulated in desiccated tissues as they would be required immediately when water becomes available and storage proteins need to be recycled. This may be physiologically comparable with the mobilization of storage reserves during seed germination.

Another type of abundant transcript encodes delta-9 desaturase catalyzing the production of unsaturated fatty acids. Such activities are correlated with increased 16:1 fatty acids that appear to inhibit spore germination of powdery mildew (*Erysiphe polygoni*) in tomato, and may be related to the release of peroxides by lipoxygenases.<sup>66</sup> Similarly, application of fatty acids in potato enhances resistance to *Erwinia coratovora* and *Phytophthora infestans*.<sup>67</sup> Delta-9 desaturase activities could also be involved in the stabilization of cell membranes and organellar structures that are required as cells rehydrate.

The GO enrichment analysis also showed that indole derivative metabolism-related transcripts were represented in the desiccated sample. These processes are related to production of indole-3-acetic acid (IAA) or to the synthesis of alkaloids known as indolalkylamines. IAA production has been reported to rapidly increase during dehydration of the resurrection plant *Craterostigma wilmsii* and also during rehydration.<sup>68</sup> These results indicate that desiccated *C. plantagineum* stores transcripts related to metabolite salvage, primary metabolism and hormones to promptly resume physiological functions. We previously proposed that enzymes of sugar metabolism are involved in such preparations for rehydration in *C. plantagineum*.<sup>69</sup>

Various other transcripts abundant during desiccation encode proteins involved in DNA integration and enzymes such as DNA-directed RNA polymerases and retrotransposon elements. The involvement of retrotransposons and siRNAs in drought tolerance has been documented. However, the mechanisms of regulation mediated by such elements remain obscure.<sup>9</sup> Lastly, desiccated *C. plantagineum* accumulates a large portion of transcripts corresponding to *Populus* and *Vitis* transcripts with unknown functions. This suggests that *C. plantagineum* shares mechanisms of desiccation tolerance with these important plants.

### The rehydration transcriptome and oxidative stress

The transcript profile during rehydration contains biotic and abiotic stress signatures and responses to oxidative species and related detoxification enzymes. Transcripts that are more abundant in rehydrated tissues encode proteins related to defense responses and oxidative stress, such that rehydration elicits responses similar to those caused by pathogens. For example, transcripts accumulate encoding cinnamoyl CoA reductase and caffeoyl-CoA O-methyltransferase involved in cell wall reinforcement elicited by pathogens.<sup>70,71</sup> These enzymes are coordinated with the phenylpropanoid pathway that provides monolignols for lignin synthesis.<sup>72</sup> Additionally, some transcripts in the rehydration profile encode enzymes related to phyloquinone metabolism. Phyloquinones (Vitamin K1) and anthraquinones are electron transfer cofactors in photosystems that are synthesized from chorismate.<sup>73,74</sup>

Various peroxidase transcripts, such as those of ascorbate peroxidases, accumulate in rehydrating samples and presumably function in detoxification with abundant glutathione- S-transferases.<sup>75</sup> However, abundant transcripts were also found encoding pathogen-responsive STH2,<sup>76</sup> PR10 and several chitinases. STH2 homologs accumulate in pea during late embryogenesis<sup>77</sup> and in birch pollen.<sup>78</sup> These proteins may play a general role in stress signaling.<sup>79</sup> A recent study identified regulatory components of an ABA receptor. These proteins, termed RCARs, are similar to PR10 and Betv1.<sup>80</sup> This suggests a more general link between biotic and abiotic stress responses mediated by ABA. Thus, the interplay between ABA sensing, oxidative stress responses and efficient re-establishment of photosynthesis may be required for rehydration and growth in *C. plantagineum*.

### Transcript profiles in growing and rehydrated leaves

Profile cluster V includes transcripts accumulating in both control and rehydrating samples. These include those of tonoplast intrinsic proteins, a subfamily of aquaporins important for water uptake and flow across membranes.<sup>81</sup> Several studies have revealed great fluctuation in the patterns of expression and tissue localization of these proteins across species. It has been proposed that such variation may be key to understanding the different strategies plants use to tolerate water stress.

A number of transcripts within this growth and rehydration cluster encode cytoskeletal proteins such as alpha tubulins. In addition, housekeeping transcripts accumulate that encode chloroplast-targeted photosystem proteins, carbohydrate metabolic enzymes and signaling components. Interestingly, GA-induced protein 3 transcripts in this cluster are not detectable in the dehydrated and desiccated samples. The expression of these proteins is inhibited by ABA treatment.<sup>82</sup> Such GA-induced cysteine-rich proteins may function in cell division and growth,<sup>83</sup> and may inhibit ABA-mediated ROS accumulation in guard cells.<sup>84</sup> Earlier investigations of desiccation tolerance in *C. plantagineum* callus indicated that ABA-GA antagonisms may be crucial in desiccation tolerance (D.B. unpublished data). Such antagonism has been well documented in seed desiccation tolerance versus germination.<sup>85</sup>

### Dehydrated and desiccated transcriptomes

Transcripts overrepresented in dehydrated and desiccated tissues form cluster VI. As expected from previous screens, LEA and desiccation-related *C. plantagineum* genes are found in this cluster. Oxidoreductase transcripts accumulate in dehydrating tissues, such as that of glyceraldehyde-3- phosphate dehydrogenase (GAPDHc), which is more abundant in dehydrated than in desiccated samples in agreement with a previous report by Velasco et al. (1994).<sup>86</sup> GAPDH enzyme activity was also found to increase during drying in Arabidopsis.<sup>87</sup> Heat and anaerobic stress also prompted its accumulation in Arabidopsis, soybean and maize.<sup>88,89</sup> GAPDH has also been proposed to have protein kinase activity, to bind RNA and even to have ribozyme and phosphotransferase activities.<sup>89</sup>

### Transcripts for key metabolic pathways are involved in desiccation tolerance in *C. plantagineum*

The GO enrichment indicates that several metabolic pathways required for desiccation tolerance are transcriptionally regulated. Although counter-intuitive, as photosynthesis and all other metabolism stops at desiccation, the regulation of photosynthesis is crucial to rehydration by priming carbon flux during water loss. Photorespiration is increased by drought to protect against photoinhibition and as a conduit for metabolites such as glycine.<sup>90</sup> In addition, carbon flux in *C. plantagineum* could be efficiently assisted by folate pathways. For instance, germinating embryos and meristematic tissues store folates as co-enzymes to re-establish metabolism and nucleotide synthesis during photoautotrophic growth.<sup>27</sup> Together with folates, the biosynthesis of cofactors such as ubiquinone and other terpenoid-quinones enhance the functions of PSI.<sup>91</sup>

In addition, lignans and cell wall materials, as well as flavonoids, isoflavonoids and other protective compounds, are readily available from the phenylpropanoid pathway. Downstream pathways are involved in geraniol degradation and the superpathway of geranylgeranyldiphosphate biosynthesis II (via methylerythritol phosphate, MEP). Our analysis of phytohormone biosynthesis-related transcripts (Figure 9.6) elucidates differential expression, especially at late steps in phytohormones such as ABA and GA synthesis. Antagonistic signaling between these hormones may finetune growth and the expression of stress responses prompted by water availability in *C. plantagineum*.

### Desiccation in *C. plantagineum* leaves and desiccation tolerance in seeds and pollen

Transcript analysis indicates that resurrection plants employ desiccation strategies commonly used by orthodox seeds and reproductive structures such as pollen. For example, thiamin metabolism is regulated transcriptionally during embryo development. Thus, the *THIC1* gene is abundantly expressed in developing maize embryos. Similarly, expression of folate biosynthetic enzymes increases during seedling growth.<sup>92</sup> Transcriptional profiling in soybean pollen grains has shown that cell wall remodeling, cell signaling and protein ubiquitination are markedly regulated throughout development.<sup>93</sup> All these processes may be related to establish the photoautotrophic phase in seedlings, or to the resumption of metabolic functions in pollen and rehydrated resurrection plants.

### Conclusions

Deep sequencing permitted us to characterize the transcript profiles of the dehydration/rehydration cycle of *C. plantagineum*. This identified hallmark genes for each hydration stage whose largely metabolic functions are required for the remarkably efficient strategies to overcome drought and hydration cycles in the field. Our analysis indicates that resurrection plants employ in vegetative tissues desiccation tolerance mechanisms similar to those of seeds and pollen. Some dehydration responsive transcripts which accumulate in desiccation-sensitive species



in response to water stress are constitutively expressed in non-stressed, control *C. plantagineum*. This is consistent with a priming strategy to account for the desiccation tolerance of resurrection plants. Furthermore, metabolic pathways controlling photosynthesis, carbon metabolism, energy, secondary metabolites and osmolyte production, as well as phytohormone regulators, were also implicated in mediating the acquisition of desiccation tolerance in *C. plantagineum*.

## Experimental procedures

### Plant growth

*Craterostigma plantagineum* (Hochst.) plants, originally collected by Professor Volk (University of Würzburg, Germany) were grown in an artificial clay substrate (Seramins; Masterfoods GmbH, <http://www.masterfoods.com>) with 16 h light (4000 lx) and 23/19°C day/ night.<sup>21</sup> Six- to eight-week-old plants were harvested or subjected to dehydration by withholding water. Early dehydrated plants (approximately 80% RWC) were collected after 48 h and the remaining plants were dehydrated for at least 21 days and reached a RWC of approximately 5%. For rehydration, dehydrated plants were removed from the substrate and submerged in water for 24 h. The RWC was determined according to Bernacchia et al. (1996).<sup>16</sup>

### RNA isolation

Total RNA was isolated according to Valenzuela-Avendano et al. (2005)<sup>94</sup> with modifications. RNA quality was gel verified and quantified spectrophotometrically (NanoDrop, ThermoScientific, <http://www.thermo.com>). Messenger RNA was isolated twice with Dynabeads Oligo (dT)<sub>25</sub> (DynaL Biotech ASA, Dynal Invitrogen, <http://www.invitrogen.com>) to minimize rRNA contamination. One microgram of mRNA per sample was used as template for first-strand cDNA synthesis using SMART technology (Clontech Laboratories Inc, <http://www.clontech.com>) to favor full-length synthesis. Double-stranded cDNA was made by 13 cycles of longdistance PCR. Complementary DNA was purified with QIAquick columns (Qiagen, <http://www.qiagen.com>) to eliminate oligo dT and enzymes. The cDNA quality was verified with an Agilent 2100 Bioanalyzer (Nimblegen, <http://www.nimblegen.com>).

### Library preparation for pyro-sequencing

Three micrograms of each cDNA sample was nebulized to produce fragments of a mean size between 400 and 800 bp. Preparation of cDNA fragment libraries and emulsion PCR were as described in the Roche GS FLX manual. Pyro-sequencing was performed on a Roche Genome Sequencer FLX instrument (Roche Diagnostics, <http://www.roche.com>).

### Cleaning and assembly of pyro-sequenced reads

The quality of reads was assessed with the SEQCLEAN EST trimming and validation tool (<http://compbio.dfci.harvard.edu/tgi/software>). Adaptor sequences used for library preparation were entered in an adaptor trimming database

in SEQCLEAN. SEQCLEAN trimpoints were merged with trimpoints of the sff file output from the 454 sequencing software, keeping the largest starting trimpoint and the smallest ending trimpoint. Trimmed reads were assembled with NEWBLER (Roche) with default parameters.

Following quality control and assembly 549 077 reads remained, mapped to 29 430 contigs. As the assembler can assign reads to more than one contig, the total number of non-unique reads mapped to contigs was 666 896. BLAST comparison of the assembled contigs to UniProt identified 15 093 UniProt entries, for which 19 311 (66%) of the contigs could be assigned to a UniProt entry. The percentage of reads that map to UniProt is constant over the four conditions before and after assembly, indicating that no bias towards certain conditions was introduced (see Table S5).

Furthermore, to estimate the degree of over-assembly due to paralogous transcripts assigned to a single UniProt entry, we identified sets of contigs that had been mapped to the same UniProt ID and then performed a BLAST between the contigs within such a set. Out of the 15 093 UniProt IDs in our assembly, 2963 had multiple contigs mapped to them. Of these, 1228, corresponding to 8.1% of our UniProt transcripts, had at least one paralogous contig-pair, where a contig-pair was defined as paralogous if the percentage of identity exceeded 40%.<sup>95</sup> This is probably an over-estimate since contig-pairs exceeding 40% identity could also exist due to splice-isoforms or imperfect assembly by the assembler.

### Contig joining and expression estimates

To account for gaps between contigs and overlapping contigs corresponding to different parts of a transcript, assembled sequences were mapped to known proteins based on similarity. The complete set of sequences in UniProt, including Swiss-Prot and TrEMBL, were used as a mapping backbone, and assembled contigs mapped to UniProt sequences using BLASTX with significance threshold  $E \leq 10^{-4}$ , and the best match for each query contig recorded.<sup>96</sup> Best hit proteins from Swiss-Prot and TrEMBL were then treated as transcript units. Their expression levels were assessed by the number of reads from all contigs mapping to a transcript. The resulting expression table consisted of read counts for contig-mapped UniProt entries versus samples.

### Expression analysis

The coefficient of variation (standard deviation/mean) was utilized as the differential expression measure across samples. A pseudo count of 3 was added to all contigs to stabilize the coefficients of variation in the lower intensity range. The data was between sample normalized by a quantile normalization procedure.<sup>97</sup> Consequently, the resulting expression levels are relative to the global expression level of the sample. The 500 transcripts with the highest coefficient of variation were considered differentially expressed, as the drop in coefficient of variation levels off after the top 500 genes (see Figure S1), and since a relatively high number of transcripts are required for the annotation analysis. Furthermore, annotation analysis is relatively robust to false positives. Partition clustering of the 500 most differentially expressed genes was performed using the Partitioning Around Medoids

(PAM) clustering algorithm.<sup>98</sup> The number of clusters was determined by increasing the number of clusters until similar expression patterns were repeated in separate clusters.

### Annotation and pathway enrichment

Gene names and species annotations for each UniProt sequence with BLASTX mapped contigs were retrieved from UniProt entries. GO annotations for UniProt identifiers were retrieved from the Gene Ontology ftp site, and parent terms then mapped using the GO.db annotation map in R. GO term enrichment among the 500 most differentially expressed genes, and among genes in PAM clusters, was performed by a hypergeometric test with Bonferroni multiple testing correction. To investigate metabolic pathways, KEGG<sup>25</sup> and MetaCyc<sup>26</sup> were considered. As the set of *C. plantagineum* UniProt transcripts had poor annotation coverage in these databases, orthologous genes in Arabidopsis were identified by BLASTP with a significance threshold of  $E \leq 10^{-4}$ . *Craterostigma plantagineum* UniProt sequences were used as queries and all Arabidopsis UniProt sequences as the database. Arabidopsis KEGG entries were extracted from ftp flat-files and mapped to UniProt identifiers. Similarly, MetaCyc was downloaded as a flat-file from BioCyc and Arabidopsis entries extracted and mapped to UniProt identifiers. Due to the paucity of annotations in the KEGG and MetaCyc databases, metabolic pathway enrichment was assessed by a test of the annotation skewness across all genes sorted on their differential expression, rather than a hypergeometric test. This was done by a Wilcoxon test followed by Bonferroni multiple testing.

Putative *A. thaliana* orthologs for the *C. plantagineum* transcripts in our assembly shown in Table 9.3 were extracted from the ‘proteome-wide’ ortholog mapping performed to retrieve KEGG annotations described above. The only exception to this is protein Q01931 in our assembly, for which we used the second top match. The reason for this is that there was no TAIR ID available and that the UniProt entry only had evidence on a transcript level and not the protein level.

Putative orthologs for 64 Medicago transcripts (7 from Boudet et al., 2006<sup>29</sup> and 57 from Buitink et al., 2006<sup>30</sup>) were identified by BLASTX using the 15 093 UniProt entries matching our *C. plantagineum* assembly, as database, with the same criteria as used for retrieving Arabidopsis orthologs.

### Sequence deposition

The complete set of 454 sequences will be deposited at GenBank upon publication. The dataset can also be obtained from the authors via FTP upon request.

### Acknowledgements

This work was supported by grants from the Danish Research Councils to JM (23030076; 272050367; 272060049; 274060460; 274060539).

### 9.3 Supplementary material

Additional Supporting Information may be found in the online version of this article:<sup>1</sup>

**Table S1.** Top 500 ranked transcripts based on their coefficient of variation across four dehydration/rehydration stages. Normalized and raw counted reads, corresponding cluster number and descriptions are shown.

**Table S2.** Gene ontology (GO) enrichment from UniProt IDs in each of the six clustered profiles. The hypergeometric *P*-value was Bonferroni adjusted. GO terms for metabolic functions, biological processes and cellular compartments are listed.

**Table S3.** Most significantly represented metabolic pathways in the transcriptomic analysis using KEGG and MetaCyc databases. The Wilcoxon *P*-value was Bonferroni adjusted.

**Table S4.** *Craterostigma plantagineum* orthologs of desiccation related genes in *Medicago* reported in the studies by Buitink et al. (2006) and Boudet et al. (2006).<sup>29,30</sup>

**Table S5.** Assembly details showing the number and percentage of reads retained after each step of the analysis. The steps include: 454 sequencing, assembly by NEWBLER (Roche), contig joining by mapping to UniProt and GO annotation.

**Figure S1.** Coefficient of variation against rank showing the cut-off of rank 500. The coefficient of variation is defined as the standard deviation divided by the mean across the four conditions.

**File S1.** All assembled transcripts annotations and expression values.

**File S2.** Mapping file for coloring metabolic pathways based on coefficient of variation.

---

<sup>1</sup><http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2010.04243.x/supinfo>



## **Part III**

# **Epilogue**



---

## Chapter 10

# Concluding remarks

---

### 10.1 Summary and perspectives

The main topic of this thesis is the identification of molecular entities that are involved in male reproductive disorders. During my PhD program I analyzed diverse data types generated by our collaborators, including transcriptomic expression and genomic variation in terms of SNPs and CNVs. Furthermore, I developed systems biology approaches to integrate such measurements with complementary information such as protein-protein interaction data, knock-out data, phenotypic information, and pathways, to provide improved candidate predictions and functional contexts.

In paper I we investigated the similarity of testicular carcinoma *in situ* cells to cells from the developing testis. The results provided further indications that non-spermatocytic testicular cancer arise due to disturbances in early testicular development. It raises the question to what extent such similarities to embryonic cells can be found in the precursor cells of other cancers, and especially whether this is more commonplace among childhood cancers than in late-onset cancers. It is however complicated by the fact that it is well-known, and particularly evident in *in vitro* cultures, that cancer cells have high genetic instability<sup>1</sup> and transform to express stem-cell like biomarkers. Similarities to embryonic cells may therefore only indicate that the precursor identified for a specific cancer is at a later stage in its development into a cancer cell. The emergence of single-cell RNA-Seq technology will help in resolving the issue of cellular heterogeneity. It also offers the opportunity to back-trace the evolution of a cancer cell population within a tumor by inferring cell population genealogy trees. Comparing such a tree with one generated by sampling a normally developing germ cell line at different time points could be of high interest to be able to determine the time-point for the emergence of CIS cells.



In paper II we analysed germline DNA from four families with testicular germ cell tumors to identify risk-associated CNVs. Given the low number of samples, family based association tests were not considered suitable, but instead we aimed to improve the confidence by placing them in a protein-protein interaction network context superimposed with database derived phenomic information. It can be noted that there were hundreds of CNVs that were recurrent in only a subset of the afflicted cases. The algorithm was able to detect a recurrent CNV at a locus with genes encoding for the relaxin peptide hormones, which is interesting due to the previous association of their receptor to cryptorchidism, and due to the known function of Relaxin 2 in pregnancy. The replication did not validate the finding, but the immunohistochemical expression analysis indicated that the protein may have a function in testis. The replication was performed on sporadic cases rather than familial cases and given the possible influence of environmental factors one may speculate that familial testicular germ cell tumors have a different etiology and are caused by a different type of genetic factors which are less prone to environmental interactions. However, two recent GWASs on testicular cancer include familial samples and the effect size estimates of the identified risk loci do not significantly differ between familial and sporadic cases.<sup>2,3</sup>

Paper III presents a genome-wide association study on testicular dysgenesis syndrome. We confirmed the importance of *KITLG* in testicular cancer and identified two risk loci at *TGFBR3* and *BMP7* by using a systems biology approach that was guided by the developmental disease hypothesis, and a pathway based analysis. Further evidence is however required for these novel findings. Apart from the analysis described in the paper, I performed analysis to search for epistatic interactions (non-additive risks) between SNPs within protein complexes. There were no indications of such interactions, but it is possible that with sequencing data and improved imputation the power for this type of analysis will improve. It would be of high interest to utilize the deeper phenotypic profiling that is available from questionnaires and electronic patient records for this cohort. One may then stratify the cohort into more specific subgroups<sup>4</sup> or scan additional phenotypic traits.<sup>5</sup> Recent successes has been reported using GWASs in conjunction with electronic medical records.<sup>6</sup> In the future such analysis may become facilitated by standardized forms at the hospitals, and genotyped cohorts may be reused as a type of "reverse genetics" approach. The project also clearly identified the need for improved matching of phenotypes between species, as parts of the mouse phenotype ontology was manually mapped to human morphologies.

Paper IV describes the analysis of copy number variants using genome-wide association data from the testicular cancer subcohort of paper III. Rare variants at *PTPN1* as well as the biological process 'regulation of cell migration' were shown to be associated to testicular germ cell tumors. Investigation of the combined effect of CNVs and SNPs could be further explored, as SNPs on a certain genetic background of CNVs could give rise to epistatic effects that are stronger than SNP-SNP interactions. For identification of germline variants in cancers it may be beneficial to create a cancer interactome based on protein interactions and somatic mutations identified in tumors and cancer cell lines. For example, the fact that *KIT* is the most frequently mutated gene in tumors<sup>7</sup> indicate that such an approach could

be successful, and further, many cancer genes are implicated in several types of malignancies.

Paper V, provides an example of the application of RNA-Seq for expression analysis of a species with an unsequenced genome. We analysed the commonly called *resurrection* plant, referring to its astonishing capability of recovering from drought. We provided the first transcriptomes of this species and by comparing unstressed with desiccated and resurrected conditions we were able to point at several pathways of interest.

In conclusion, this thesis contributes to the molecular understanding of testicular malfunction and desiccation tolerance in *C. plantagineum*, as well as develops and highlights the usefulness of novel systems biology methodologies.

## 10.2 Outlook

Nearly 800 genome-wide association studies covering more than 150 distinct human diseases and traits have been published, with approximately 900 SNP-trait associations reported as significant ( $P < 5 * 10^{-8}$ , NHGRI GWAS catalog, Jan. 2011<sup>8</sup>). This has yielded insight into unexpected involvement of certain functional and mechanistic pathways in a variety of disease processes. Despite this, these variants have been of limited value to predict the individual patient's level of risk for a particular condition.<sup>9</sup> Extending the phenotypic screening to a large number of diseases may prove to provide a higher clinical value, and the genetic profile could become useful in the selection of personalized medicine when the influence of genetic variants on drug response has been more completely mapped.

The missing heritability of common diseases remaining after the pursue of genome-wide association studies has lead to an intense debate of its possible whereabouts, and has started to shift the focus of the community towards other factors, including structural variants, epigenetic influences, genetic background effects such as epistasis, and rare variants.<sup>10,11</sup> One interesting population genetics argument for the importance of rare variants is that one of the effects of the explosive human population growth that has occurred within the last 10,000 years is that it gives rise to an allele frequency spectrum strongly shifted towards a high amount of rare alleles as compared to a neutral coalescent model. This hypothesis is corroborated by a recent study that resequenced two diabetes-associated genes in more than 13,000 individuals, which enabled the possibility to estimate the amount of very rare variants.<sup>12</sup> This, and the advent of commonplace sequencing, raises the need for computational and statistical tools to assess the effect of rare variants.<sup>13</sup>

The influence of environmental factors on the development of disease is far from understood, and it presents a challenging task.<sup>14</sup> An interesting aspect of this is the effect of host-virus or host-bacteria interactions, and this has started to be addressed by the meta-sequencing projects.<sup>15</sup> In general, a trend towards the sampling of multiple biological features from the same samples should be beneficial, such as expression, genotyping and chromatin modification, as well as environmental or lifestyle profiling. This exacerbates the need of systems biology computational models.

Relatedly, condition specific experimental data, such as the evaluation of spatiotemporal differences, or the effect of systematic perturbation of systems by stresses or chemicals is demanded. For instance, the National Institutes of Health recently launched an initiative to systematically map tissue specific eQTLs (the Genotype-Tissue Expression (GTEx) project).<sup>16</sup> Such studies may also help to connect the germline variations identified by GWAS to the correct tissue, as a risk-associated genetic variant is not necessarily directly affecting the disease afflicted tissue, but it could mediate an indirect disturbance via signaling from other cell-types.

Finally, the increasing wealth of high-throughput data necessitate persistent high computation power growth rates and invites data technology and infrastructures that facilitate increased formalization and standardization. Technological advances will continue to push the border of science, and one can only speculate where the advent of stem cell technology, biological nanotechnology and synthetic biology eventually will lead us.

---

# Bibliography for Chapter 1

---

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* 458, 719–724 (2009).
2. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science (New York, N.Y.)* 322, 881–888 (2008).

---

## Bibliography for Chapter 2

---

1. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
2. Antonarakis, S. E., Chakravarti, A., Cohen, J. C. & Hardy, J. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nature reviews. Genetics* **11**, 380–384 (2010).
3. Online Mendelian Inheritance In Man. URL <http://www.ncbi.nlm.nih.gov/omim/>.
4. Thompson, E. A. R.A. Fisher’s contributions to genetical statistics. *Biometrics* **46**, 905–914 (1990).
5. Plomin, R., Haworth, C. M. A. & Davis, O. S. P. Common disorders are quantitative traits. *Nature Reviews Genetics* **10**, 872–878 (2009).
6. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
7. Falconer, D. S. & Mackay, T. F. C. *Introduction to quantitative genetics* (Longman, 1996).
8. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)* **273**, 1516–1517 (1996).
9. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
10. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308–311 (2001).
11. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36**, 949–951 (2004).
12. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
13. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
14. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
15. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
16. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
17. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

18. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2009).
19. Bray, F. *et al.* Trends in testicular cancer incidence and mortality in 22 European countries: Continuing increases in incidence and declines in mortality. *International Journal of Cancer* **118**, 3099–3111 (2006).
20. Carlsen, E., Giwercman, A., Keiding, N. & Skakkebaek, N. E. Evidence for decreasing quality of semen during past 50 years. *BMJ (Clinical research ed.)* **305**, 609–613 (1992).
21. Toppari, J., Kaleva, M. & Virtanen, H. E. Trends in the incidence of cryptorchidism and hypospadias, and methodological limitations of registry-based data. *Human reproduction update* **7**, 282–286 (2001).
22. Garner, M. J., Turner, M. C., Ghadirian, P. & Krewski, D. Epidemiology of testicular cancer: an overview. *International journal of cancer. Journal international du cancer* **116**, 331–339 (2005).
23. Sharpe, R. M. & Skakkebaek, N. E. Are oestrogens involved in falling sperm counts and disorders of the male reproductive tract? *Lancet* **341**, 1392–1395 (1993).
24. Rajpert-De Meyts, E. Developmental model for the pathogenesis of testicular carcinoma in situ: genetic and environmental aspects. *Human reproduction update* **12**, 303–323 (2006).
25. Skakkebaek, N. E., Rajpert-De Meyts, E. & Main, K. M. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects: Opinion. *Human Reproduction* **16**, 972–978 (2001).
26. Hemminki, K., Sundquist, J. & Lorenzo Bermejo, J. Familial Risks for Cancer as the Basis for Evidence-Based Clinical Referral and Counseling. *Oncologist* **13**, 239–247 (2008).
27. Swerdlow, A. J., De Stavola, B. L., Swanwick, M. A. & Maconochie, N. E. Risks of breast and testicular cancers in young adult twins in England and Wales: evidence on prenatal and genetic aetiology. *Lancet* **350**, 1723–1728 (1997).
28. Chia, V. M. *et al.* International trends in the incidence of testicular cancer, 1973–2002. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **19**, 1151–1159 (2010).
29. Crockford, G. P. *et al.* Genome-wide linkage screen for testicular germ cell tumour susceptibility loci. *Hum. Mol. Genet.* **15**, 443–451 (2006).
30. Nathanson, K. L. *et al.* The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *American journal of human genetics* **77**, 1034–1043 (2005).
31. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807–810 (2009).
32. Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811–815 (2009).
33. Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature genetics* **42**, 604–607 (2010).
34. Turnbull, C. *et al.* UK Genome Wide Association Study in Testicular Germ Cell Tumor. *Speaker abstract, 7th Copenhagen Workshop on CIS - Testis and Germ Cell Cancer* (2010).
35. Rapley, E. A. & Nathanson, K. L. Predisposition alleles for testicular germ cell tumour. *Current Opinion in Genetics & Development* **20**, 225–230 (2010).
36. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–D950 (2011).
37. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. *Int. J. Cancer* **99**, 260–266 (2002).

38. Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. *Nature Medicine* **14**, 1197–1213 (2008).
39. Main, K. M., Skakkebaek, N. E., Virtanen, H. E. & Toppari, J. Genital anomalies in boys and the environment. *Best Practice & Research Clinical Endocrinology & Metabolism* **24**, 279–289 (2010).
40. Yhee, J. H. & Baskin, L. S. Environmental factors in genitourinary development. *The Journal of urology* **184**, 34–41 (2010).
41. Chia, V. M. *et al.* Effect modification of endocrine disruptors and testicular germ cell tumour risk by hormone-metabolizing genes. *International journal of andrology* **33**, 588–596 (2010).

---

## Bibliography for Chapter 3

---

1. Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *American journal of human genetics* **69**, 936–950 (2001).
2. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Human molecular genetics* **11**, 2417–2423 (2002).
3. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics* **6**, 95–108 (2005).
4. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends in genetics : TIG* **17**, 502–510 (2001).
5. Neel, J. V. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *American journal of human genetics* **14**, 353–362 (1962).
6. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
7. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* **308**, 385–389 (2005).
8. Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–9367 (2009).
9. Gibson, G. The environmental contribution to gene expression profiles. *Nature reviews. Genetics* **9**, 575–581 (2008).
10. Jensen, L. J., Jensen, T. S., de Lichtenberg, U., Brunak, S. & Bork, P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443**, 594–597 (2006).
11. Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415–1426 (2007).
12. Gavin, A.-C. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
13. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science (New York, N.Y.)* **303**, 540–543 (2004).
14. Giot, L. *et al.* A Protein Interaction Map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
15. Rual, J.-F. F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).



16. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Molecular systems biology* 5 (2009).
17. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics (Oxford, England)* 26, 1057–1063 (2010).
18. Wu, X., Liu, Q. & Jiang, R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics (Oxford, England)* 25, 98–104 (2009).
19. Wu, X., Jiang, R., Zhang, M. Q. & Li, S. Network-based global inference of human disease genes. *Molecular systems biology* 4 (2008).
20. Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences of the United States of America* 105, 4323–4328 (2008).
21. Park, J., Lee, D.-S. S., Christakis, N. A. & Barabási, A.-L. L. The impact of cellular networks on disease comorbidity. *Molecular systems biology* 5 (2009).
22. Goh, K.-I. I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* 104, 8685–8690 (2007).
23. Chavali, S., Barrenas, F., Kanduri, K. & Benson, M. Network properties of human disease genes with pleiotropic effects. *BMC systems biology* 4, 78+ (2010).
24. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nature methods* 6, 91–97 (2009).
25. Cusick, M. E. *et al.* Literature-curated protein interaction datasets. *Nature methods* 6, 39–46 (2009).
26. Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nature reviews. Genetics* 11, 259–272 (2010).
27. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* 375, 1525–1535 (2010).
28. Galvan, A., Ioannidis, J. P. & Dragani, T. A. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in genetics : TIG* 26, 132–141 (2010).
29. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* 4, 682–690 (2008).
30. Yildirim, M. A., Goh, K.-I. I., Cusick, M. E., Barabási, A.-L. L. & Vidal, M. Drug-target network. *Nature biotechnology* 25, 1119–1126 (2007).
31. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* 406, 378–382 (2000).
32. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391 (2002).
33. Taboureau, O. *et al.* ChemProt: a disease chemical biology database. *Nucleic acids research* 39, D367–D372 (2011).
34. Vidal, M. A unifying view of 21st century systems biology. *FEBS letters* 583, 3891–3894 (2009).
35. Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America* 43, 553–566 (1957).
36. Monod, J. & Jacob, F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor symposia on quantitative biology* 26, 389–401 (1961).
37. Mahadevan, R., Palsson, B. & Lovley, D. R. In situ to in silico and back: elucidating the physiology and ecology of *Geobacter* spp. using genome-scale modelling. *Nature reviews. Microbiology* 9, 39–50 (2011).

38. Keller, P. J., Schmidt, A. D., Wittbrodt, J. & Stelzer, E. H. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science (New York, N.Y.)* **322**, 1065–1069 (2008).
39. Quigley, D. & Balmain, A. Systems genetics analysis of cancer susceptibility: from mouse models to humans. *Nature reviews. Genetics* **10**, 651–657 (2009).
40. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
41. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009).
42. Li, M.-X. X., Sham, P. C., Cherny, S. S. & Song, Y.-Q. Q. A knowledge-based weighting framework to boost the power of genome-wide association studies. *PLoS one* **5**, e14480+ (2010).

---

## Bibliography for Chapter 4

---

1. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 5, R80+ (2004).
2. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 3 (2004).
3. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 98, 5116–5121 (2001).
4. McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Human Molecular Genetics* 17, R135–R142 (2008).
5. Slager, S. L. & Schaid, D. J. Case-control studies of genetic markers: power and sample size approximations for Armitage’s test for trend. *Human heredity* 52, 149–153 (2001).
6. Zheng, G., Freidlin, B. & Gastwirth, J. L. Comparison of robust tests for genetic association using case-control studies. vol. 49, 253–265 (Institute of Mathematical Statistics, 2006).
7. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics* 9, 356–369 (2008).
8. Scharpf, R. B., Parmigiani, G., Pevsner, J. & Ruczinski, I. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *The annals of applied statistics* 2, 687–713 (2008).
9. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* 40, 1253–1260 (2008).
10. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research* 35, 2013–2025 (2007).
11. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* 17, 1665–1674 (2007).
12. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459, 987–991 (2009).
13. Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nature genetics* 40, 1245–1252 (2008).
14. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* 11, 843–854 (2010).

15. Hakes, L., Robertson, D. L., Oliver, S. G. & Lovell, S. C. Protein interactions from complexes: a structural perspective. *Comparative and functional genomics* (2007).
16. Zanzoni, A. *et al.* MINT: a Molecular INteraction database. *FEBS letters* **513**, 135–140 (2002).
17. Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic acids research* **31**, 248–250 (2003).
18. Breitkreutz, B.-J. J., Stark, C. & Tyers, M. The GRID: the General Repository for Interaction Datasets. *Genome Biol* **4** (2003).
19. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic acids research* **32** (2004).
20. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic acids research* **32** (2004).
21. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic acids research* **28**, 289–291 (2000).
22. Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P. & Weinstein, H. PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics (Oxford, England)* **21**, 827–828 (2005).
23. D’Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Methods in molecular biology (Clifton, N.J.)* **694**, 49–61 (2011).
24. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**, 29–34 (1999).
25. Shou, C. *et al.* Measuring the Evolutionary Rewiring of Biological Networks. *PLoS Comput Biol* **7**, e1001050+ (2011).
26. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nature methods* **6**, 91–97 (2009).
27. Scott, M. S. & Barton, G. J. Probabilistic prediction and ranking of human protein-protein interactions. *BMC bioinformatics* **8**, 239+ (2007).
28. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
29. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science (New York, N.Y.)* **270**, 484–487 (1995).
30. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* **18**, 630–634 (2000).
31. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**, 1509–1517 (2008).
32. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature protocols* **5**, 516–535 (2010).
33. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature biotechnology* **27**, 455–457 (2009).
34. Oshlack, A. & Wakefield, M. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct* **4**, 14+ (2009).
35. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis* (Wiley, 1990).
36. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561–D568 (2011).

37. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic acids research* 37, D674–679 (2009).
38. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25–29 (2000).
39. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* 39, D945–D950 (2011).
40. Gauthier, N. P., Jensen, L. J., Wernersson, R., Brunak, S. & Jensen, T. S. Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Research* 38, D699–D702 (2010).
41. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* 25, 309–316 (2007).
42. Online Mendelian Inheritance In Man. URL <http://www.ncbi.nlm.nih.gov/omim/>.
43. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23, 657–663 (2007).

---

## Bibliography for Chapter 5

---

1. Rørth, M. *et al.* Carcinoma in situ in the testis. *Scandinavian journal of urology and nephrology. Supplementum* 166–186 (2000).
2. Skakkebaek, N. E., Berthelsen, J. G. & Müller, J. Carcinoma-in-situ of the undescended testis. *The Urologic clinics of North America* 9, 377–385 (1982).
3. Skakkebaek, N. E., Rajpert-De Meyts, E. & Main, K. M. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects. *Human reproduction (Oxford, England)* 16, 972–978 (2001).
4. Sonne, S. B. *et al.* Testicular dysgenesis syndrome and the origin of carcinoma in situ testis. *International journal of andrology* 31, 275–287 (2008).
5. Holstein, A. F. & Körner, F. Light and electron microscopical analysis of cell types in human seminoma. *Virchows Archiv. A, Pathological anatomy and histology* 363, 97–112 (1974).
6. Rajpert-De Meyts, E. *et al.* The emerging phenotype of the testicular carcinoma in situ germ cell. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* 111 (2003).
7. Almstrup, K. *et al.* Embryonic stem cell-like features of testicular carcinoma in situ revealed by genome-wide gene expression profiling. *Cancer research* 64, 4736–4743 (2004).
8. Draper, J. S. *et al.* Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells. *Nature biotechnology* 22, 53–54 (2004).
9. Kraggerud, S. M. M. *et al.* Genome profiles of familial/bilateral and sporadic testicular germ cell tumors. *Genes, chromosomes & cancer* 34, 168–174 (2002).
10. Rosenberg, C., Van Gorp, R. J., Geelen, E., Oosterhuis, J. W. & Looijenga, L. H. Overrepresentation of the short arm of chromosome 12 is related to invasive growth of human testicular seminomas and nonseminomas. *Oncogene* 19, 5858–5862 (2000).
11. Adamah, D. J. *et al.* Dysfunction of the mitotic/meiotic switch as a potential cause of neoplastic conversion of primordial germ cells. *International journal of andrology* 29, 219–227 (2006).
12. Biermann, K. *et al.* Gene expression profiling identifies new biological markers of neoplastic germ cells. *Anticancer research* 27, 3091–3100 (2007).
13. Skotheim, R. I. *et al.* Differentiation of human embryonal carcinomas in vitro and in vivo reveals expression profiles relevant to normal development. *Cancer research* 65, 5588–5598 (2005).
14. Sonne, S. B. B. *et al.* Optimizing staining protocols for laser microdissection of specific cell types from the testis including carcinoma in situ. *PloS one* 4 (2009).

15. Høeie-Hansen, C. E. *et al.* Transcription factor AP-2gamma is a developmentally regulated marker of testicular carcinoma in situ and germ cell tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research* **10**, 8521–8530 (2004).
16. Jensen, C. H., Erb, K., Westergaard, L. G., Kliem, A. & Teisner, B. Fetal antigen 1, an EGF multidomain protein in the sex hormone-producing cells of the gonads and the microenvironment of germ cells. *Molecular human reproduction* **5**, 908–913 (1999).
17. Rajpert-De Meyts, E. *et al.* Expression of anti-Müllerian hormone during normal and pathological gonadal development: association with differentiation of Sertoli and granulosa cells. *The Journal of clinical endocrinology and metabolism* **84**, 3836–3844 (1999).
18. Visfeldt, J., Cortes, D., Thorup, J. M. & Byskov, A. G. Anti-MIC2 as a tool in examination of testicular biopsies. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* **107**, 631–635 (1999).
19. Saeed, A. I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **34**, 374–378 (2003).
20. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121 (2001).
21. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S. Fourth Edition* (Springer, 2002).
22. Nielsen, J. E. *et al.* Germ cell differentiation-dependent and stage-specific expression of LANCL1 in rodent testis. *European journal of histochemistry : EJH* **47**, 215–222 (2003).
23. Rajpert-De Meyts, E. *et al.* Developmental expression of POU5F1 (OCT-3/4) in normal and dysgenetic human gonads. *Human reproduction (Oxford, England)* **19**, 1338–1344 (2004).
24. Miettinen, M., Virtanen, I. & Talerma, A. Intermediate filament proteins in human testis and testicular germ-cell tumors. *The American journal of pathology* **120**, 402–410 (1985).
25. Czernobilsky, B., Moll, R., Levy, R. & Franke, W. W. Co-expression of cytokeratin and vimentin filaments in mesothelial, granulosa and rete ovarii cells of the human ovary. *European journal of cell biology* **37**, 175–190 (1985).
26. O’Shaughnessy, P. J. *et al.* Developmental changes in human fetal testicular cell numbers and messenger ribonucleic acid levels during the second trimester. *The Journal of clinical endocrinology and metabolism* **92**, 4792–4801 (2007).
27. Hersmus, R. *et al.* FOXL2 and SOX9 as parameters of female and male gonadal differentiation in patients with various forms of disorders of sex development (DSD). *The Journal of pathology* **215**, 31–38 (2008).
28. Steger, K., Pauls, K., Klönisch, T., Franke, F. E. & Bergmann, M. Expression of protamine-1 and -2 mRNA during human spermiogenesis. *Molecular human reproduction* **6**, 219–225 (2000).
29. Sonne, S. B. B. *et al.* Identity of M2A (D2-40) antigen and gp36 (Aggrus, T1A-2, podoplanin) in human developing testis, testicular carcinoma in situ and germ-cell tumours. *Virchows Archiv : an international journal of pathology* **449**, 200–206 (2006).
30. Yeung, K. Y. Y. & Bumgarner, R. E. Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome biology* **4** (2003).
31. Radhakrishnan, Y. *et al.* Identification, characterization, and evolution of a primate beta-defensin gene cluster. *Genes and immunity* **6**, 203–210 (2005).
32. Berger, F., Lau, C., Dahlmann, M. & Ziegler, M. Subcellular compartmentation and differential catalytic properties of the three human nicotinamide mononucleotide adenylyltransferase isoforms. *The Journal of biological chemistry* **280**, 36334–36341 (2005).
33. Goldmann, T., Otto, F. & Vollmer, E. A receptor-type protein tyrosine phosphatase PTP zeta is expressed in human cutaneous melanomas. *Folia histochemica et cytobiologica / Polish Academy of Sciences, Polish Histochemical and Cytochemical Society* **38**, 19–20 (2000).

34. Katoh, M. & Katoh, M. Identification and characterization of ASXL3 gene in silico. *International journal of oncology* **24**, 1617–1622 (2004).
35. Cools, M., Drop, S. L., Wolffenbuttel, K. P., Oosterhuis, J. W. & Looijenga, L. H. Germ cell tumors in the intersex gonad: old paths, new directions, moving frontiers. *Endocrine reviews* **27**, 468–484 (2006).
36. Giwercman, A., Marks, A., Bailey, D., Bauml, R. & Skakkebaek, N. E. A monoclonal antibody as a marker for carcinoma in situ germ cells of the human adult testis. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* **96**, 667–670 (1988).
37. Jørgensen, N. *et al.* Expression of immunohistochemical markers for testicular carcinoma in situ by normal human fetal germ cells. *Laboratory investigation; a journal of technical methods and pathology* **72**, 223–231 (1995).
38. Jacobsen, G. K. & Nørgaard-Pedersen, B. Placental alkaline phosphatase in testicular germ cell tumours and in carcinoma-in-situ of the testis. An immunohistochemical study. *Acta pathologica, microbiologica, et immunologica Scandinavica. Section A, Pathology* **92**, 323–329 (1984).
39. Hoei-Hansen, C. E. *et al.* Stem cell pluripotency factor NANOG is expressed in human fetal gonocytes, testicular carcinoma in situ and germ cell tumours. *Histopathology* **47**, 48–56 (2005).
40. Looijenga, L. H. *et al.* POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors. *Cancer research* **63**, 2244–2250 (2003).
41. Honecker, F. *et al.* Pathobiological implications of the expression of markers of testicular carcinoma in situ by fetal germ cells. *The Journal of pathology* **203**, 849–857 (2004).
42. Clark, A. T. The stem cell identity of testicular cancer. *Stem cell reviews* **3**, 49–59 (2007).
43. Conrad, S. *et al.* Generation of pluripotent stem cells from adult human testis. *Nature* **456**, 344–349 (2008).
44. Skakkebaek, N. E. Possible carcinoma-in-situ of the testis. *Lancet* **2**, 516–517 (1972).
45. Müller, J., Skakkeboek, N. E. & Lundsteen, C. Aneuploidy as a marker for carcinoma-in-situ of the testis. *Acta pathologica et microbiologica Scandinavica. Section A, Pathology* **89**, 67–68 (1981).
46. Chaganti, R. S. & Houldsworth, J. Genetics and biology of adult human male germ cell tumors. *Cancer research* **60**, 1475–1482 (2000).
47. Ottesen, A. M. M. *et al.* High-resolution comparative genomic hybridization detects extra chromosome arm 12p material in most cases of carcinoma in situ adjacent to overt germ cell tumors, but not before the invasive tumor development. *Genes, chromosomes & cancer* **38**, 117–125 (2003).
48. Looijenga, L. H. *et al.* Role of gain of 12p in germ cell tumour development. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* **111** (2003).
49. Chemes, H. *et al.* Early manifestations of testicular dysgenesis in children: pathological phenotypes, karyotype correlations and precursor stages of tumour development. *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica* **111** (2003).
50. Sharpe, R. M. & Skakkebaek, N. E. Testicular dysgenesis syndrome: mechanistic insights and potential new downstream effects. *Fertility and sterility* **89** (2008).



---

## Bibliography for Chapter 6

---

1. Bray, F. *et al.* Trends in testicular cancer incidence and mortality in 22 European countries: Continuing increases in incidence and declines in mortality. *International Journal of Cancer* **118**, 3099–3111 (2006).
2. Hemminki, K., Sundquist, J. & Lorenzo Bermejo, J. Familial Risks for Cancer as the Basis for Evidence-Based Clinical Referral and Counseling. *Oncologist* **13**, 239–247 (2008).
3. Crockford, G. P. *et al.* Genome-wide linkage screen for testicular germ cell tumour susceptibility loci. *Hum. Mol. Genet.* **15**, 443–451 (2006).
4. Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811–815 (2009).
5. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807–810 (2009).
6. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
7. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
8. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
9. Fanciulli, M. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics* **39**, 721–723 (2007).
10. Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
11. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25**, 309–316 (2007).
12. Kirchhoff, M., Bisgaard, A., Duno, M., Hansen, F. & Schwartz, M. A 17q21.31 microduplication, reciprocal to the newly described 17q21.31 microdeletion, in a girl with severe psychomotor developmental delay and dysmorphic craniofacial features. *European Journal of Medical Genetics* **50**, 256–263 (2007).
13. Venkatraman, E. S. & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657–663 (2007).
14. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature genetics* **36**, 949–951 (2004).

15. Ottesen, A. M., Garn, I. D., Akglaede, L., Juul, A. & Rajpert-De Meyts, E. A simple screening method for detection of Klinefelter syndrome and other X-chromosome aneuploidies based on copy number of the androgen receptor gene. *Mol. Hum. Reprod.* **13**, 745–750 (2007).
16. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology (Clifton, N.J.)* **132**, 365–386 (2000).
17. Mau Kai, C. *et al.* Sons conceived by assisted reproduction techniques inherit deletions in the azoospermia factor (AZF) region of the Y chromosome and the DAZ gene copy number. *Human reproduction (Oxford, England)* **23**, 1669–1678 (2008).
18. Sokol, R. Z., Wang, X. S., Lechago, J., Johnston, P. D. & Swerdloff, R. S. Immunohistochemical localization of relaxin in human prostate. *The journal of histochemistry and cytochemistry* **37**, 1253–1255 (1989).
19. Hansell, D. J., Bryant-Greenwood, G. D. & Greenwood, F. C. Expression of the Human Relaxin H1 Gene in the Decidua, Trophoblast, and Prostate. *J Clin Endocrinol Metab* **72**, 899–904 (1991).
20. Feng, S. *et al.* Relaxin promotes prostate cancer progression. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 1695–1702 (2007).
21. Ottesen, A. M. *et al.* Cytogenetic and molecular analysis of a family with three brothers afflicted with germ-cell cancer. *Clinical genetics* **65**, 32–39 (2004).
22. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
23. van der Zwaag, B. *et al.* Gene-Network Analysis Identifies Susceptibility Genes Related to Glycobiology in Autism. *PLoS ONE* **4**, e5324+ (2009).
24. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
25. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
26. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**, 1665–1674 (2007).
27. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**, 1166–1174 (2008).
28. Shaikh, T. H. *et al.* High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome research* **19**, 1682–1690 (2009).
29. Bailey, J. A. *et al.* Recent Segmental Duplications in the Human Genome. *Science* **297**, 1003–1007 (2002).
30. Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nature Genetics* **39**, S22–S29 (2007).
31. Van Der Westhuizen, E. T., Summers, R. J., Halls, M. L., Bathgate, R. A. & Sexton, P. M. Relaxin receptors—new drug targets for multiple disease states. *Current drug targets* **8**, 91–104 (2007).
32. Ivell, R. & Grutzner, F. Evolution and male fertility: lessons from the insulin-like factor 6 gene (Insl6). *Endocrinology* **150**, 3986–3990 (2009).
33. Olesen, I. A. *et al.* Testicular carcinoma in situ in subfertile Danish men. *International journal of andrology* **30**, 406–412 (2007).
34. Petersen, P. M., Skakkebaek, N. E., Vistisen, K., Rørth, M. & Giwercman, A. Semen quality and reproductive hormones before orchiectomy in men with testicular cancer. *Journal of clinical oncology* **17**, 941–947 (1999).

35. Skakkebaek, N. E., Rajpert-De Meyts, E. & Main, K. M. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects: Opinion. *Human Reproduction* **16**, 972–978 (2001).
36. Ivell, R., Kotula-Balak, M., Glynn, D., Heng, K. & Anand-Ivell, R. Relaxin family peptides in the male reproductive system - a critical appraisal. *Molecular human reproduction* (2010).
37. Yki-Järvinen, H., Wahlström, T. & Seppälä, M. Immunohistochemical demonstration of relaxin in the genital tract of men. *Journal of reproduction and fertility* **69**, 693–695 (1983).
38. Lessing, J. B. *et al.* Effect of relaxin on human spermatozoa. *The Journal of reproductive medicine* **31**, 304–309 (1986).
39. Kohsaka, T. *et al.* Identification of boar testis as a source and target tissue of relaxin. *Annals of the New York Academy of Sciences* **1160**, 194–196 (2009).
40. Cardoso, L. C., Nascimento, A. R., Royer, C., Porto, C. S. & Lazari, M. F. Locally produced relaxin may affect testis and vas deferens function in rats. *Reproduction (Cambridge, England)* **139**, 185–196 (2010).
41. Sonne, S. B. *et al.* Analysis of Gene Expression Profiles of Microdissected Cell Populations Indicates that Testicular Carcinoma In situ Is an Arrested Gonocyte. *Cancer Res* **69**, 5241–5250 (2009).
42. Filonzi, M. *et al.* Relaxin family peptide receptors Rxfp1 and Rxfp2: mapping of the mRNA and protein distribution in the reproductive tract of the male rat. *Reproductive biology and endocrinology : RB&E* **5**, 29+ (2007).
43. Samuel, C. S., Tian, H., Zhao, L. & Amento, E. P. Relaxin is a key mediator of prostate growth and male reproductive tract development. *Laboratory investigation; a journal of technical methods and pathology* **83**, 1055–1067 (2003).

---

## Bibliography for Chapter 7

---

1. Schnack, T. H., Poulsen, G., Myrup, C., Wohlfahrt, J. & Melbye, M. Familial coaggregation of cryptorchidism, hypospadias, and testicular germ cell cancer: a nationwide cohort study. *Journal of the National Cancer Institute* **102**, 187–192 (2010).
2. Skakkebaek, N. E., Rajpert-De Meyts, E. & Main, K. M. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects: Opinion. *Human Reproduction* **16**, 972–978 (2001).
3. Chia, V. M. *et al.* International trends in the incidence of testicular cancer, 1973–2002. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **19**, 1151–1159 (2010).
4. Engholm, G. *et al.* NORDCAN—a Nordic tool for cancer information, planning, quality control and research. *Acta oncologica (Stockholm, Sweden)* **49**, 725–736 (2010).
5. Asklund, C. *et al.* Semen quality, reproductive hormones and fertility of men operated for hypospadias. *International journal of andrology* **33**, 80–87 (2010).
6. Rajpert-De Meyts, E. Developmental model for the pathogenesis of testicular carcinoma in situ: genetic and environmental aspects. *Human reproduction update* **12**, 303–323 (2006).
7. Sonne, S. B. *et al.* Analysis of Gene Expression Profiles of Microdissected Cell Populations Indicates that Testicular Carcinoma In situ Is an Arrested Gonocyte. *Cancer Res* **69**, 5241–5250 (2009).
8. Skakkebaek, N. E., Berthelsen, J. G., Giwercman, A. & Müller, J. Carcinoma-in-situ of the testis: possible origin from gonocytes and precursor of all types of germ cell tumours except spermatocytoma. *International journal of andrology* **10**, 19–28 (1987).
9. Sharpe, R. M. & Skakkebaek, N. E. Testicular dysgenesis syndrome: mechanistic insights and potential new downstream effects. *Fertility and sterility* **89** (2008).
10. Boisen, K. A. *et al.* Difference in prevalence of congenital cryptorchidism in infants between two Nordic countries. *Lancet* **363**, 1264–1269 (2004).
11. Jørgensen, N. *et al.* East-West gradient in semen quality in the Nordic-Baltic area: a study of men from the general population in Denmark, Norway, Estonia and Finland. *Human reproduction (Oxford, England)* **17**, 2199–2208 (2002).
12. Hemminki, K. & Li, X. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. *British journal of cancer* **90**, 1765–1770 (2004).
13. Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811–815 (2009).
14. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807–810 (2009).

15. Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature genetics* **42**, 604–607 (2010).
16. Zhong, H., Yang, X., Kaplan, L. M., Molony, C. & Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *American journal of human genetics* **86**, 581–591 (2010).
17. Naukkarinen, J. *et al.* Use of Genome-Wide Expression Data to Mine the "Gray Zone" of GWA Studies Leads to Novel Candidate Obesity Genes. *PLoS Genet* **6**, e1000976+ (2010).
18. Hsu, Y.-H. H. *et al.* An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. *PLoS genetics* **6**, e1000977+ (2010).
19. Pers, T. H., Hansen, N. T., Lage, K., Koefoed, P. & Dworzynski, P. Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genetic Epidemiology* **In press** (2011).
20. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* **11**, 843–854 (2010).
21. Li, Q., Zheng, G., Li, Z. & Yu, K. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. *Annals of human genetics* **72**, 397–406 (2008).
22. Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405–417 (2003).
23. Small, C. L., Shima, J. E., Uzumcu, M., Skinner, M. K. & Griswold, M. D. Profiling Gene Expression During the Differentiation and Development of the Murine Embryonic Gonad. *Biology of Reproduction* **72**, 492–501 (2005).
24. McMahon, A. P. *et al.* GUDMAP: The Genitourinary Developmental Molecular Anatomy Project. *J Am Soc Nephrol* **19**, 667–671 (2008).
25. Houmard, B. *et al.* Global gene expression in the human fetal testis and ovary. *Biology of reproduction* **81**, 438–443 (2009).
26. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25**, 309–316 (2007).
27. Thorup, J. *et al.* What is new in cryptorchidism and hypospadias—a critical review on the testicular dysgenesis hypothesis. *Journal of pediatric surgery* **45**, 2074–2086 (2010).
28. Kondo, T., Zákány, J., Innis, J. W. & Duboule, D. Of fingers, toes and penises. *Nature* **390** (1997).
29. Wang, Y. *et al.* Allelic variants in HOX genes in cryptorchidism. *Birth defects research. Part A, Clinical and molecular teratology* **79**, 269–275 (2007).
30. Gonzalez, F., Duboule, D. & Spitz, F. Transgenic analysis of Hoxd gene regulation during digit development. *Developmental biology* **306**, 847–859 (2007).
31. Goodman, F. R., Majewski, F., Collins, A. L. & Scambler, P. J. A 117-kb microdeletion removing HOXD9–HOXD13 and EVX2 causes synpolydactyly. *American journal of human genetics* **70**, 547–555 (2002).
32. Galan, J. J. *et al.* Association of genetic markers within the KIT and KITLG genes with human male infertility. *Human reproduction (Oxford, England)* **21**, 3185–3192 (2006).
33. Dias, V. L., Rajpert-De Meyts, E., McLachlan, R. & Loveland, K. L. L. Analysis of activin/TGF $\beta$ -signaling modulators within the normal and dysfunctional adult human testis reveals evidence of altered signaling capacity in a subset of seminomas. *Reproduction (Cambridge, England)* **138**, 801–811 (2009).
34. Sarraj, M. A. *et al.* Fetal testis dysgenesis and compromised Leydig cell function in Tgfb3 (beta glycan) knockout mice. *Biology of reproduction* **82**, 153–162 (2010).

35. Anderson, R. A., Cambray, N., Hartley, P. S. & McNeilly, A. S. Expression and localization of inhibin alpha, inhibin/activin betaA and betaB and the activin type II and inhibin beta-glycan receptors in the developing human testis. *Reproduction (Cambridge, England)* **123**, 779–788 (2002).
36. Cobellis, L. *et al.* Gonadal malignant germ cell tumors express immunoreactive inhibin/activin subunits. *European journal of endocrinology / European Federation of Endocrine Societies* **145**, 779–784 (2001).
37. Dias, V. *et al.* Activin receptor subunits in normal and dysfunctional adult human testis. *Human reproduction (Oxford, England)* **23**, 412–420 (2008).
38. Skakkebaek, N. E., Rajpert-De Meyts, E. & Main, K. M. Testicular dysgenesis syndrome: an increasingly common developmental disorder with environmental aspects. *Human reproduction (Oxford, England)* **16**, 972–978 (2001).
39. Galwey, N. W. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic epidemiology* **33**, 559–568 (2009).
40. Sidak, Z. On Multivariate Normal Probabilities of Rectangles: Their Dependence on Correlations. *The Annals of Mathematical Statistics* **39** (1968).
41. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology* **3** (2002).
42. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S233–S240 (2002).

---

## Bibliography for Chapter 8

---

1. Chia, V. M. *et al.* International trends in the incidence of testicular cancer, 1973-2002. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **19**, 1151-1159 (2010).
2. Swerdlow, A. J., De Stavola, B. L., Swanwick, M. A. & Maconochie, N. E. Risks of breast and testicular cancers in young adult twins in England and Wales: evidence on prenatal and genetic aetiology. *Lancet* **350**, 1723-1728 (1997).
3. Hemminki, K. & Li, X. Familial risk in testicular cancer as a clue to a heritable and environmental aetiology. *British journal of cancer* **90**, 1765-1770 (2004).
4. Nathanson, K. L. *et al.* The Y deletion gr/gr and susceptibility to testicular germ cell tumor. *American journal of human genetics* **77**, 1034-1043 (2005).
5. Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811-815 (2009).
6. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807-810 (2009).
7. Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature genetics* **42**, 604-607 (2010).
8. Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987-991 (2009).
9. Fanciulli, M. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics* **39**, 721-723 (2007).
10. Hollox, E. J. *et al.* Psoriasis is associated with increased  $\beta$ -defensin genomic copy number. *Nature Genetics* **40**, 23-25 (2007).
11. Yang, T.-L. L. *et al.* Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *American journal of human genetics* **83**, 663-674 (2008).
12. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
13. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237-241 (2008).
14. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).
15. Glessner, J. T. *et al.* A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *American journal of human genetics* **87**, 661-666 (2010).

16. Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nature genetics* **40**, 1245–1252 (2008).
17. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
18. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics (Oxford, England)* **26**, 1057–1063 (2010).
19. Efron, B. & Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology* **23**, 70–86 (2002).
20. Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
21. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
22. Gamazon, E. R., Nicolae, D. L. & Cox, N. J. A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. *PLoS genetics* **7** (2011).
23. Lessard, L., Stuiblé, M. & Tremblay, M. L. The two faces of PTP1B in cancer. *Biochimica et biophysica acta* **1804**, 613–619 (2010).
24. Nishizaki, T. *et al.* Investigation of genetic alterations associated with the grade of astrocytic tumor by comparative genomic hybridization. *Genes Chromosom. Cancer* **21**, 340–346 (1998).
25. Schaid, D. J. The complex genetic epidemiology of prostate cancer. *Human molecular genetics* **13 Spec No 1** (2004).
26. Furukawa, T., Sunamura, M. & Horii, A. Molecular mechanisms of pancreatic carcinogenesis. *Cancer science* **97**, 1–7 (2006).
27. Tanner, M. M. *et al.* Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. *Cancer research* **56**, 3441–3445 (1996).
28. Stuiblé, M., Doody, K. M. & Tremblay, M. L. PTP1B and TC-PTP: regulators of transformation and tumorigenesis. *Cancer metastasis reviews* **27**, 215–230 (2008).
29. Brasil, A. S. *et al.* PTPN11 and KRAS gene analysis in patients with Noonan and Noonan-like syndromes. *Genetic testing and molecular biomarkers* **14**, 425–432 (2010).
30. Freeman, A. K. & Monteiro, A. N. Phosphatases in the cellular response to DNA damage. *Cell communication and signaling : CCS* **8**, 27+ (2010).
31. Rapley, E. A. & Nathanson, K. L. Predisposition alleles for testicular germ cell tumour. *Current Opinion in Genetics & Development* **20**, 225–230 (2010).
32. Matzuk, M. M. & Lamb, D. J. Genetic dissection of mammalian fertility pathways. *Nature cell biology* **4 Suppl** (2002).
33. Oosterhuis, J. W. & Looijenga, L. H. J. Testicular germ-cell tumours in a broader perspective. *Nature Reviews Cancer* **5**, 210–222 (2005).
34. Xu, Y. *et al.* Receptor type protein tyrosine phosphatase-kappa mediates cross-talk between transforming growth factor-beta and epidermal growth factor receptor signaling pathways in human keratinocytes. *Molecular biology of the cell* **21**, 29–35 (2010).
35. Lawson, K. A. *et al.* Bmp4 is required for the generation of primordial germ cells in the mouse embryo. *Genes & development* **13**, 424–436 (1999).
36. Ying, Y., Liu, X. M., Marble, A., Lawson, K. A. & Zhao, G. Q. Requirement of Bmp8b for the generation of primordial germ cells in the mouse. *Molecular endocrinology (Baltimore, Md.)* **14**, 1053–1063 (2000).



37. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253–1260 (2008).
38. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**, 29–34 (1999).
39. D’Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Methods in molecular biology (Clifton, N.J.)* **694**, 49–61 (2011).
40. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic acids research* **37**, D674–679 (2009).
41. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25–29 (2000).
42. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–D950 (2011).
43. Gauthier, N. P., Jensen, L. J., Wernersson, R., Brunak, S. & Jensen, T. S. Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Research* **38**, D699–D702 (2010).
44. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* **25**, 309–316 (2007).
45. Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. *Nature Medicine* **14**, 1197–1213 (2008).

---

## Bibliography for Chapter 9

---

1. Gaff, D. F. Desiccation-tolerant flowering plants in southern Africa. *Science (New York, N.Y.)* **174**, 1033–1034 (1971).
2. Porembski, S. & Barthlott, W. Granitic and gneissic outcrops (inselbergs) as centers of diversity for desiccation-tolerant vascular plants. *Plant Ecology* **151**, 19–28 (2000).
3. Proctor, M. C. & Pence, V. C. *Vegetative Tissues: Bryophytes, Vascular Plants and Vegetative Propagules* (CABI Publishing, Wallingford, UK, 2002).
4. Oliver, M. J., Velten, J. & Mishler, B. D. Desiccation Tolerance in Bryophytes: A Reflection of the Primitive Strategy for Plant Survival in Dehydrating Habitats? *Integrative and Comparative Biology* **45**, 788–799 (2005).
5. Ingram, J. & Bartels, D. The molecular basis of dehydration tolerance in plants. *Annual review of plant physiology and plant molecular biology* **47**, 377–403 (1996).
6. Oliver, M. J., Tuba, Z. & Mishler, B. D. The evolution of vegetative desiccation tolerance in land plants. *Plant Ecology* **151**, 85–100 (2000).
7. Sherwin, H. W. & Farrant, J. M. Protection mechanisms against excess light in the resurrection plants *Craterostigma wilmsii* and *Xerophyta viscosa*. *Plant Growth Regulation* **24**, 203–210 (1998).
8. Furini, A., Koncz, C., Salamini, F. & Bartels, D. High level transcription of a member of a repeated gene family confers dehydration tolerance to callus tissue of *Craterostigma plantagineum*. *EMBO J* **16**, 3599–3608 (1997).
9. Hilbricht, T. *et al.* Retrotransposons and siRNA have a role in the evolution of desiccation tolerance leading to resurrection of the plant *Craterostigma plantagineum*. *New Phytologist* **179**, 877–887 (2008).
10. Bianchi, G., Gamba, A., Murelli, C., Salamini, F. & Bartels, D. Novel carbohydrate metabolism in the resurrection plant *Craterostigma plantagineum*. *Plant Journal* **1**, 355–359 (1991).
11. Norwood, M., Truesdale, M. R., Richter, A. & Scott, P. Photosynthetic carbohydrate metabolism in the resurrection plant *Craterostigma plantagineum*. *Journal of experimental botany* **51**, 159–165 (2000).
12. Bartels, D. Desiccation Tolerance Studied in the Resurrection Plant *Craterostigma plantagineum*. *Integrative and Comparative Biology* **45**, 696–701 (2005).
13. Hoekstra, F. A., Golovina, E. A. & Buitink, J. Mechanisms of plant desiccation tolerance. *Trends in plant science* **6**, 431–438 (2001).
14. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).

15. Tunc-Ozdemir, M. *et al.* Thiamin confers enhanced tolerance to oxidative stress in Arabidopsis. *Plant physiology* **151**, 421–432 (2009).
16. Bernacchia, G., Salamini, F. & Bartels, D. Molecular Characterization of the Rehydration Process in the Resurrection Plant *Craterostigma plantagineum*. *Plant physiology* **111**, 1043–1050 (1996).
17. Bockel, C., Salamini, F. & Bartels, D. Isolation and characterization of genes expressed during early events of the dehydration process in the resurrection plant *Craterostigma plantagineum*. *Journal of Plant Physiology* **152**, 158–166 (1998).
18. Bartels, D. & Salamini, F. Desiccation tolerance in the resurrection plant *Craterostigma plantagineum*. A contribution to the study of drought tolerance at the molecular level. *Plant physiology* **127**, 1346–1353 (2001).
19. Velasco, R., Salamini, F. & Bartels, D. Gene structure and expression analysis of the drought- and abscisic acid-responsive CDeT11-24 gene family from the resurrection plant *Craterostigma plantagineum* Hochst. *Planta* **204**, 459–471 (1998).
20. Alamillo, J. M., Roncarati, R. & Heino, P. Molecular analysis of desiccation tolerance in barley embryos and in the resurrection plant *Craterostigma plantagineum*. *Agronomie* **2**, 161–167 (1994).
21. Bartels, D., Schneider, K., Terstappen, G., Piatkowski, D. & Salamini, F. Molecular cloning of abscisic acid-modulated genes which are induced during desiccation of the resurrection plant *Craterostigma plantagineum*. *Planta* **181**, 27–34 (1990).
22. Kirch, H.-H., Nair, A. & Bartels, D. Novel ABA- and dehydration-inducible aldehyde dehydrogenase genes isolated from the resurrection plant *Craterostigma plantagineum* and Arabidopsis thaliana. *The Plant Journal* **28**, 555–567 (2001).
23. Frank, W., Munnik, T., Kerkmann, K., Salamini, F. & Bartels, D. Water deficit triggers phospholipase D activity in the resurrection plant *Craterostigma plantagineum*. *The Plant cell* **12**, 111–124 (2000).
24. Bernacchia, G., Schwall, G., Lottspeich, F., Salamini, F. & Bartels, D. The transketolase gene family of the resurrection plant *Craterostigma plantagineum*: differential expression during the rehydration phase. *The EMBO journal* **14**, 610–618 (1995).
25. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**, 29–34 (1999).
26. Karp, P. D., Riley, M., Paley, S. M. & Pellegrini-Toole, A. The MetaCyc Database. *Nucleic acids research* **30**, 59–61 (2002).
27. Jabrin, S., Ravel, S., Gambonnet, B., Douce, R. & Rébeillé, F. One-carbon metabolism in plants. Regulation of tetrahydrofolate synthesis during germination and seedling development. *Plant physiology* **131**, 1431–1439 (2003).
28. Kilian, J. *et al.* The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal* **50**, 347–363 (2007).
29. Boudet, J. *et al.* Comparative analysis of the heat stable proteome of radicles of *Medicago truncatula* seeds during germination identifies late embryogenesis abundant proteins associated with desiccation tolerance. *Plant Physiol* (2006).
30. Buitink, J. *et al.* Transcriptome profiling uncovers metabolic and regulatory processes occurring during the transition from desiccation-sensitive to desiccation-tolerant stages in *Medicago truncatula* seeds. *The Plant Journal* **47**, 735–750 (2006).
31. Urano, K. *et al.* Characterization of the ABA-regulated global responses to dehydration in Arabidopsis by metabolomics. *The Plant Journal* **57**, 1065–1078 (2009).
32. Cramer, G. R. *et al.* Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Functional & integrative genomics* **7**, 111–134 (2007).

33. Wood, A. J. & Oliver, M. J. Translational control in plant stress: the formation of messenger ribonucleoprotein particles (mRNPs) in response to desiccation of *Tortula ruralis* gametophytes. *The Plant Journal* **18**, 359–370 (1999).
34. Oliver, M. J., Dowd, S. E., Zaragoza, J., Mauget, S. A. & Payton, P. R. The rehydration transcriptome of the desiccation-tolerant bryophyte *Tortula ruralis*: transcript classification and analysis. *BMC Genomics* **5** (2004).
35. Iturriaga, G., Cushman, M. & Cushman, J. An EST catalogue from the resurrection plant *Selaginella lepidophylla* reveals abiotic stress-adaptive genes. *Plant Science* **170**, 1173–1184 (2006).
36. Neale, A. D. *et al.* The isolation of genes from the resurrection grass *Sporobolus stapfianus* which are induced during severe drought stress. *Plant, Cell & Environment* **23**, 265–277 (2000).
37. Taji, T. *et al.* Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *The Plant Journal* **29**, 417–426 (2002).
38. Saravitz, D. M., Pharr, D. M. & Carter, T. E. Galactinol synthase activity and soluble sugars in developing seeds of four soybean genotypes. *Plant physiology* **83**, 185–189 (1987).
39. Koster, K. L. & Leopold, A. C. Sugars and desiccation tolerance in seeds. *Plant physiology* **88**, 829–832 (1988).
40. Downie, B. *et al.* Expression of a GALACTINOL SYNTHASE gene in tomato seeds is up-regulated before maturation desiccation and again after imbibition whenever radicle protrusion is prevented. *Plant physiology* **131**, 1347–1359 (2003).
41. Hammond, J. P. & White, P. J. Sucrose transport in the phloem: integrating root responses to phosphorus starvation. *Journal of experimental botany* **59**, 93–109 (2008).
42. Wu, P. *et al.* Phosphate starvation triggers distinct alterations of genome expression in *Arabidopsis* roots and leaves. *Plant physiology* **132**, 1260–1271 (2003).
43. Cosgrove, D. J. Enzymes and other agents that enhance cell wall extensibility. *Annual review of plant physiology and plant molecular biology* **50**, 391–417 (1999).
44. Jones, L. & McQueen-Mason, S. A role for expansins in dehydration and rehydration of the resurrection plant *Craterostigma plantagineum*. *FEBS letters* **559**, 61–65 (2004).
45. Vicré, M., Lerouxel, O., Farrant, J., Lerouge, P. & Driouich, A. Composition and desiccation-induced alterations of the cell wall in the resurrection plant *Craterostigma wilmsii*. *Physiologia Plantarum* **120**, 229–239 (2004).
46. Ahn, I.-P. P., Kim, S. & Lee, Y.-H. H. Vitamin B1 functions as an activator of plant disease resistance. *Plant physiology* **138**, 1505–1515 (2005).
47. Ahn, I.-P. P., Kim, S., Lee, Y.-H. H. & Suh, S.-C. C. Vitamin B1-induced priming is dependent on hydrogen peroxide and the NPR1 gene in *Arabidopsis*. *Plant physiology* **143**, 838–848 (2007).
48. Roje, S. Vitamin B biosynthesis in plants. *Phytochemistry* **68**, 1904–1921 (2007).
49. Settembre, E., Begley, T. P. & Ealick, S. E. Structural biology of enzymes of the thiamin biosynthesis pathway. *Current opinion in structural biology* **13**, 739–747 (2003).
50. Ribeiro, D. T. T. *et al.* Functional characterization of the *thi1* promoter region from *Arabidopsis thaliana*. *Journal of experimental botany* **56**, 1797–1804 (2005).
51. Rapala-Kozik, M., Kowalska, E. & Ostrowska, K. Modulation of thiamine metabolism in *Zea mays* seedlings under conditions of abiotic stress. *Journal of experimental botany* **59**, 4133–4143 (2008).
52. Zhu, J.-K. K. Salt and drought stress signal transduction in plants. *Annual review of plant biology* **53**, 247–273 (2002).
53. Conley, T. R., Peng, H. P. & Shih, M. C. Mutations affecting induction of glycolytic and fermentative genes during germination and environmental stresses in *Arabidopsis*. *Plant physiology* **119**, 599–608 (1999).

54. Kürsteiner, O., Dupuis, I. & Kuhlemeier, C. The pyruvate decarboxylase1 gene of Arabidopsis is required during anoxia but not other environmental stresses. *Plant physiology* **132**, 968–978 (2003).
55. Schenk, G., Layfield, R., Candy, J. M., Duggleby, R. G. & Nixon, P. F. Molecular evolutionary analysis of the thiamine-diphosphate-dependent enzyme, transketolase. *Journal of molecular evolution* **44**, 552–572 (1997).
56. Willige, B. C., Kutzer, M., Tebartz, F. & Bartels, D. Subcellular localization and enzymatic properties of differentially expressed transketolase genes isolated from the desiccation tolerant resurrection plant *Craterostigma plantagineum*. *Planta* **229**, 659–666 (2009).
57. Charron, J. F. & Sarhan, F. *Plant lipocalins*. In *Lipocalins*, 41–48 (Landes Bioscience, Georgetown, TX, USA, 2006).
58. Demel, R. A. & De Kruyff, B. The function of sterols in membranes. *Biochimica et biophysica acta* **457**, 109–132 (1976).
59. Arnoux, P., Morosinotto, T., Saga, G., Bassi, R. & Pignol, D. A structural basis for the pH-dependent xanthophyll cycle in Arabidopsis thaliana. *The Plant cell* **21**, 2036–2044 (2009).
60. Breton, G., Danyluk, J., Charron, J.-B. F. B. & Sarhan, F. Expression profiling and bioinformatic analyses of a novel stress-regulated multispinning transmembrane protein family from cereals and Arabidopsis. *Plant physiology* **132**, 64–74 (2003).
61. Medina, J., Catalá, R. & Salinas, J. Developmental and stress regulation of RCI2A and RCI2B, two cold-inducible genes of Arabidopsis encoding highly conserved hydrophobic proteins. *Plant physiology* **125**, 1655–1666 (2001).
62. Hu, H. *et al.* Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12987–12992 (2006).
63. Gosti, F. *et al.* ABI1 protein phosphatase 2C is a negative regulator of abscisic acid signaling. *The Plant cell* **11**, 1897–1910 (1999).
64. Solomon, M., Belenghi, B., Delledonne, M., Menachem, E. & Levine, A. The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *The Plant cell* **11**, 431–444 (1999).
65. Watanabe, H., Abe, K., Emori, Y., Hosoyama, H. & Arai, S. Molecular cloning and gibberellin-induced expression of multiple cysteine proteinases of rice seeds (oryzains). *The Journal of biological chemistry* **266**, 16897–16902 (1991).
66. Wang, C. Expression of the yeast  $\Delta$ -9 desaturase gene in tomato enhances its resistance to powdery mildew. *Physiological and Molecular Plant Pathology* **52**, 371–383 (1998).
67. Cohen, Y., Gisi, U. & Mosinger, E. Systemic resistance of potato plants against *Phytophthora infestans* induced by unsaturated fatty acids. *Physiological and Molecular Plant Pathology* **38**, 255–263 (1991).
68. Vicré, M., Farrant, J. M. & Driouich, A. Insights into the cellular mechanisms of desiccation tolerance among angiosperm resurrection plant species. *Plant, Cell & Environment* **27**, 1329–1340 (2004).
69. Kleines, M. *et al.* Isolation and expression analysis of two stress-responsive sucrose-synthase genes from the resurrection plant *Craterostigma plantagineum* (Hochst.). *Planta* **209**, 13–24 (1999).
70. Schmitt, D., Pakusch, A. E. & Matern, U. Molecular cloning, induction and taxonomic distribution of caffeoyl-CoA 3-O-methyltransferase, an enzyme involved in disease resistance. *The Journal of biological chemistry* **266**, 17416–17423 (1991).

71. Tronchet, M., Balagué, C., Kroj, T., Jouanin, L. & Roby, D. Cinnamyl alcohol dehydrogenases-C and D, key enzymes in lignin biosynthesis, play an essential role in disease resistance in *Arabidopsis*. *Molecular Plant Pathology* **11**, 83–92 (2010).
72. Maury, S., Geoffroy, P. & Legrand, M. Tobacco O-methyltransferases involved in phenylpropanoid metabolism. The different caffeoyl-coenzyme A/5-hydroxyferuloyl-coenzyme A 3/5-O-methyltransferase and caffeic acid/5-hydroxyferulic acid 3/5-O-methyltransferase classes have distinct substrate specificities and expression patterns. *Plant physiology* **121**, 215–224 (1999).
73. Poulsen, C. Roles of chorismate mutase, isochorismate synthase and anthranilate synthase in plants. *Phytochemistry* **30**, 377–386 (1991).
74. Shimada, H. *et al.* Inactivation and deficiency of core proteins of photosystems I and II caused by genetical phyloquinone and plastoquinone deficiency but retained lamellar structure in a T-DNA mutant of *Arabidopsis*. *The Plant Journal* **41**, 627–637 (2005).
75. Dixon, D., Laphorn, A. & Edwards, R. Plant glutathione transferases. *Genome Biology* **3** (2002).
76. Matton, D. P., Prescott, G., Bertrand, C., Camirand, A. & Brisson, N. Identification of cis-acting elements involved in the regulation of the pathogenesis-related gene STH-2 in potato. *Plant molecular biology* **22**, 279–291 (1993).
77. Barratt, D. H. P. & Clark, J. A. Proteins arising during the late stages of embryogenesis in *Pisum sativum*. *Planta* **184**, 14–23 (1991).
78. Swoboda, I. *et al.* Bet v 1 proteins, the major birch pollen allergens and members of a family of conserved pathogenesis-related proteins, show ribonuclease activity in vitro. *Physiologia Plantarum* **96**, 433–438 (1996).
79. Pühringer, H. The promoter of an apple Ypr10 gene, encoding the major allergen Mal d 1, is stress- and pathogen-inducible. *Plant Science* **152**, 35–50 (2000).
80. Ma, Y. *et al.* Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science (New York, N.Y.)* **324**, 1064–1068 (2009).
81. Alexandersson, E. *et al.* Whole gene family expression and drought stress regulation of aquaporins. *Plant molecular biology* **59**, 469–484 (2005).
82. Shi, L. & Olszewski, N. E. Gibberellin and abscisic acid regulate GAST1 expression at the level of transcription. *Plant molecular biology* **38**, 1053–1060 (1998).
83. Ben-Nissan, G., Lee, J.-Y. Y., Borohov, A. & Weiss, D. GIP, a *Petunia hybrida* GA-induced cysteine-rich protein: a possible role in shoot elongation and transition to flowering. *The Plant journal : for cell and molecular biology* **37**, 229–238 (2004).
84. Wigoda, N., Ben-Nissan, G., Granot, D., Schwartz, A. & Weiss, D. The gibberellin-induced, cysteine-rich protein GIP2 from *Petunia hybrida* exhibits in planta antioxidant activity. *The Plant Journal* **48**, 796–805 (2006).
85. Seo, M. *et al.* Regulation of hormone metabolism in *Arabidopsis* seeds: phytochrome regulation of abscisic acid metabolism and abscisic acid regulation of gibberellin metabolism. *The Plant Journal* **48**, 354–366 (2006).
86. Velasco, R., Salamini, F. & Bartels, D. Dehydration and ABA increase mRNA levels and enzyme activity of cytosolic GAPDH in the resurrection plant *Craterostigma plantagineum*. *Plant molecular biology* **26**, 541–546 (1994).
87. Gallardo, K. *et al.* Proteomic analysis of *Arabidopsis* seed germination and priming. *Plant physiology* **126**, 835–848 (2001).
88. Russell, D. A., Wong, D. M. & Sachs, M. M. The anaerobic response of soybean. *Plant physiology* **92**, 401–407 (1990).
89. Chang, W. W. *et al.* Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment, and identification of proteins by mass spectrometry. *Plant physiology* **122**, 295–318 (2000).

90. Takahashi, S., Bauwe, H. & Badger, M. Impairment of the photorespiratory pathway accelerates photoinhibition of photosystem II by suppression of repair but not acceleration of damage processes in Arabidopsis. *Plant physiology* **144**, 487–494 (2007).
91. Gross, J. *et al.* A plant locus essential for phylloquinone (vitamin K1) biosynthesis originated from a fusion of four eubacterial genes. *The Journal of biological chemistry* **281**, 17189–17196 (2006).
92. Belanger, F. C., Leustek, T., Chu, B. & Kriz, A. L. Evidence for the thiamine biosynthetic pathway in higher-plant plastids and its developmental regulation. *Plant molecular biology* **29**, 809–821 (1995).
93. Haerizadeh, F., Wong, C. E., Bhalla, P. L., Gresshoff, P. M. & Singh, M. B. Genomic expression profiling of mature soybean (*Glycine max*) pollen. *BMC plant biology* **9**, 25+ (2009).
94. Valenzuela-Avendaño, J. *et al.* Use of a simple method to isolate intact RNA from partially hydrated *Selaginella lepidophylla* plants. *Plant Molecular Biology Reporter* **23**, 199–200 (2005).
95. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16**, 1667–1678 (2004).
96. Meyer, E. *et al.* Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFLx. *BMC genomics* **10**, 219+ (2009).
97. Workman, C. *et al.* A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology* **3** (2002).
98. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis* (Wiley, 1990).

---

# Bibliography for Chapter 10

---

1. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
2. Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807–810 (2009).
3. Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811–815 (2009).
4. Plomin, R., Haworth, C. M. A. & Davis, O. S. P. Common disorders are quantitative traits. *Nature Reviews Genetics* **10**, 872–878 (2009).
5. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)* **26**, 1205–1210 (2010).
6. Ritchie, M. D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *American journal of human genetics* **86**, 560–572 (2010).
7. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, D945–D950 (2011).
8. Hindorf, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9362–9367 (2009).
9. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *The New England journal of medicine* **363**, 166–176 (2010).
10. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**, 446–450 (2010).
11. Heard, E. *et al.* Ten years of genetics and genomics: what have we achieved and where are we heading? *Nature Reviews Genetics* **11**, 723–733 (2010).
12. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature communications* **1**, 131+ (2010).
13. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature reviews. Genetics* **11**, 773–785 (2010).
14. Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nature reviews. Genetics* **11**, 259–272 (2010).
15. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
16. The NIH Common Fund's Genotype-Tissue Expression (GTEx) program. URL <http://commonfund.nih.gov/GTEx/>.



---

## Appendix A. Supplementary material paper II

---

Supplementary Table 1

**Supplementary Table 1**

Specification of investigated families and samples

Family	Individual	Germ cell tumor	244K chip	185K chip	RLN1-qPCR validation study
1	brother 1	yes	x	x	x
	brother 2	yes	x	x	x
	brother 3	yes	x	x	x
	mother	no	x	x	x
	father	no	x	x	x
	sister	no	x	-	x
2	brother 1	yes	-	x	x
	brother 2	yes	-	x	-
	mother	no	-	x	-
	son of brother 1	no	-	x	-
3	son	yes	x	-	x
	father	yes	x	-	x
4	brother 1	yes	x	-	x
	brother 2	yes	x	-	x
	mother	no	x	-	-
	uncle 1 (maternal)	no	x	-	x
	uncle 2 (maternal)	no	-	-	x
	cousin	no	-	-	x
5	brother 1	yes	-	-	x
	brother 2	no	-	-	-
	father	yes	-	-	-
6	son	yes	-	-	x
	father	yes	-	-	x
	cousin (paternal)	yes	-	-	-

Supplementary Table 2

Supplementary Table 2

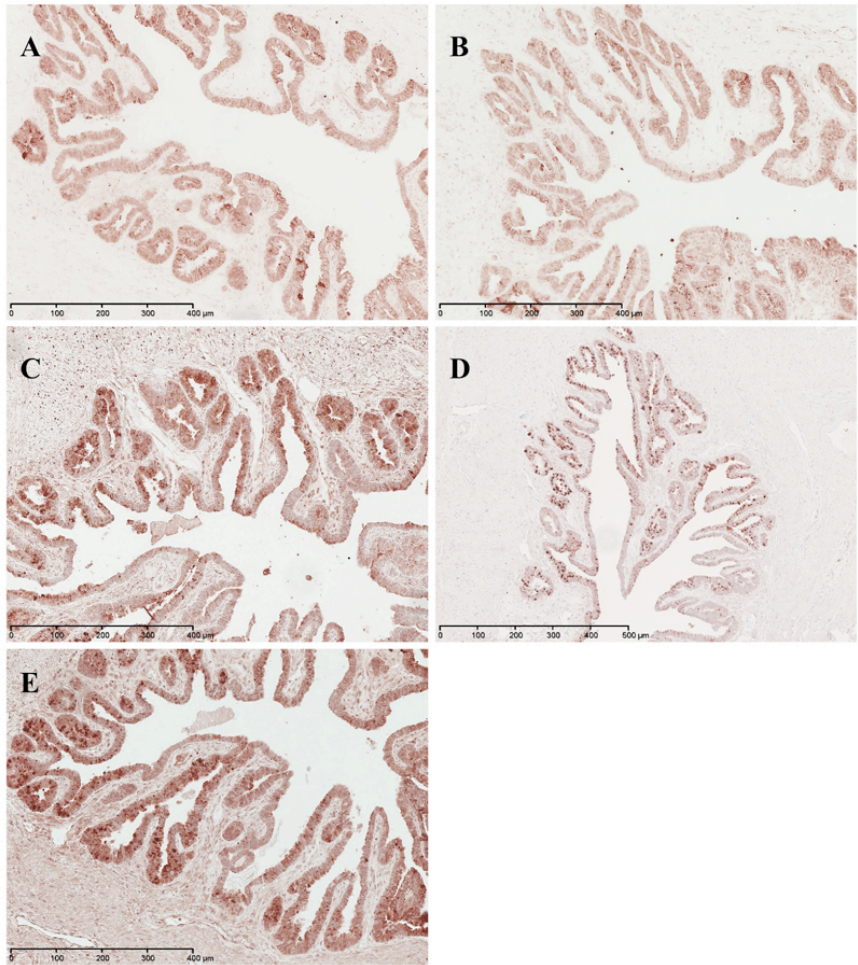
Recurrent CNVs occurring in all GCT afflicted individuals within a family, and CNVs found in at least five individuals from at least two different families. All GCT cases are marked in red. The four CNVs marked in yellow indicate loci that are not common in the Database of Genomic Variants.

Family	CNVs recurring in all three GCT afflicted brothers	Type	Gene relation	DGV	Mother	Father	Brother			Sister
							1	2	3	
1	chr6:32567382..32600962	Loss	<i>HLA-DRB5</i> exon	Common	1	1	1	1	1	1
	chr7:142159154..142171618	Loss	25kb upstream <i>PRSS2</i> , exon of ESTs	Common	0	1	1	1	1	1
	chr9:5298149..5325065	Loss	Exon of <i>RLNI</i>	Rare (Shaikh 2009: 1.0%, McCarroll 2008: 1.5%)	1	0	1	1	1	0
	chr14:19268776..19490630	Loss	Olfactory receptors	Common	0	1	1	1	1	0
	chr15:19786655..20079935	Loss	Olfactory receptors	Common	1	0	1	1	1	0
	chr16:33205648..33539023	Loss	300kb downsteraam of <i>TP53TG3</i>	Common	0	1	1	1	1	1
	chr16:62882631..63119309	Gain	<i>CDH11</i> 500kb downstream	Rare (one sample in DGV)	0	1	1	1	1	1
	chr8:7100835..7885555	Loss	<i>DEFB</i> genes and <i>SPAG11B</i>	Common	1	1	1	1	1	N
2							Brother 1	Brother 2	Mother	Son of Brother 1
							1	2		1
	chr1:149369523..149391633	Loss	<i>SEMA6C</i> , exons	Rare (no CNV in DGV)	1	1	0	0		
	chr6:32539864..32654018	Loss	<i>HLA-DRB5</i> exon	Common	1	1	1	0		
	chr8:39356595..39499752	Loss	<i>ADAMP5P</i> , tMDC	Common	1	1	1	1		
	chr8:7233982..7885555	Loss	<i>DEFB</i> genes and <i>SPAG11B</i>	Common	1	1	1	0		
	chr14:18642252..19490630	Loss	Olfactory receptors	Common	1	1	1	1		
	chr16:32105104..33536799	Loss	300kb downsteraam of <i>TP53TG3</i>	Common	1	1	0	1		
	chr22:37683612..37709939	Loss	Between <i>APOBEC3A</i> and <i>APOBEC3B</i>	Common	1	1	1	0		
	chr11:55124799..55206590	Gain	Olfactory receptors	Common	1	1	0	0		
3							Father	Son		
							1	1		
	chr3:163997228..164101776	Loss	EST <i>BC073807</i> , 2Mb from <i>B3GALNT1</i>	Common	1	1				
	chr6:211079..320890	Gain	<i>DUSP22</i>	Common	1	1				
	chr8:39356595..39505256	Gain	<i>ADAMP5P</i> , tMDC	Common	1	1				
	chr10:46404919..46568496	Gain	<i>GPRIN2</i> , <i>PPYR1</i> exons	Common	1	1				
chrY:22505342..24754207	Gain	<i>DAZ1</i> , <i>DAZ2</i> , <i>DAZ4</i> region	Common	1	1					

Supplementary Table 2

	Loss in at least 5 samples and from at least two families	n loss	Gene relation	DGV	Fam.1:	Fam.1:	Fam.1:	Fam.1:	Fam.1:	Fam.1:	Fam.2:	Fam.2:	Fam.2:	Fam.2:	
					Mother	Father	Brother 1	Brother 2	Brother 3	Sister	Brother 1	Brother 2	Mother	Son of case 1	
Joint analysis of all four families	chr6:32567382..32600962	12	<i>HLA-DRB5</i> exon	Common	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	
	chr14:19446048..19490630	9	Olfactory receptors	Common	0	-1	-1	-1	-1	0	-1	-1	-1	-1	
	chr16:33280759..33536799	8	300kb downsternam of <i>TP53TG3</i>	Common	0	-1	-1	-1	-1	-1	-1	-1	0	-1	
	chr8:7729311..7789937	7	<i>DEFB</i> genes and <i>SPAG11B</i>	Common	-1	-1	-1	0	0	-1	-1	-1	-1	0	
	chr15:19886840..20060061	7	Olfactory receptors	Common	-1	0	-1	-1	-1	0	-1	1	1	0	
	chr3:3.163997228..164101776	6	EST <i>BC073807</i> , 2Mb from <i>B3GALNT1</i>	Common	-1	-1	1	0	0	-1	1	0	1	0	
	chr14:105966378..105994646	5	antibody region	Common	0	-1	-1	0	-1	-1	0	0	0	-1	
	chr17:41578267..41706870	5	<i>KIAA1267</i>	Common	-1	0	0	0	0	-1	-1	0	0	0	
	<b>Gain in at least 5 samples and from at least two families</b>		<b>n gain</b>	<b>Gene relation</b>	<b>DGV</b>										
	chr11:55124799..55194990	8	Olfactory receptors	Common	1	-1	0	1	1	1	1	1	0	-1	
chr8:39356595..39505256	7	<i>ADAMP5P</i> , tMDC	Common	1	0	1	0	0	1	-1	-1	-1	-1		
chr6:252939..320890	6	<i>DUSP22</i>	Common	1	-1	0	0	0	0	-1	0	-1	-1		
chr4:69069563..69165814	6	<i>UGT2B17</i>	Common	-1	1	1	-1	0	-1	0	0	1	1		
chr9:138516970..138534059	5	<i>NOTCH1</i> exons	Rare (2-2.5%)	1	0	0	1	0	1	0	0	0	0		

	Loss in at least 5 samples and from at least two families	n loss	Gene relation	DGV	Fam.3:	Fam.3:	Fam.4:	Fam.4:	Fam.4:	Fam.4:	
					Father	Son	Brother 1	Brother 2	Uncle	Mother	
Joint analysis of all four families (cont.)	chr6:32567382..32600962	12	<i>HLA-DRB5</i> exon	Common	1	-1	-1	0	-1	1	
	chr14:19446048..19490630	9	Olfactory receptors	Common	0	-1	1	-1	1	0	
	chr16:33280759..33536799	8	300kb downsternam of <i>TP53TG3</i>	Common	0	1	0	1	0	0	
	chr8:7729311..7789937	7	<i>DEFB</i> genes and <i>SPAG11B</i>	Common	0	0	0	1	1	1	
	chr15:19886840..20060061	7	Olfactory receptors	Common	0	-1	1	0	0	-1	
	chr3:3.163997228..164101776	6	EST <i>BC073807</i> , 2Mb from <i>B3GALNT1</i>	Common	-1	-1	-1	1	0	0	
	chr14:105966378..105994646	5	antibody region	Common	0	0	0	0	0	0	
	chr17:41578267..41706870	5	<i>KIAA1267</i>	Common	0	0	0	-1	0	-1	
	<b>Gain in at least 5 samples and from at least two families</b>		<b>n gain</b>	<b>Gene relation</b>	<b>DGV</b>						
	chr11:55124799..55194990	8	Olfactory receptors	Common	0	0	-1	1	0	1	
chr8:39356595..39505256	7	<i>ADAMP5P</i> , tMDC	Common	1	1	0	1	1	0		
chr6:252939..320890	6	<i>DUSP22</i>	Common	1	1	1	0	1	1		
chr4:69069563..69165814	6	<i>UGT2B17</i>	Common	0	0	0	1	0	1		
chr9:138516970..138534059	5	<i>NOTCH1</i> exons	Rare (2-2.5%)	0	0	1	1	0	0		



**Figure 1** – Immunohistochemical localization of relaxins and RXFP1 protein in prostate tissue, which was used as a positive control in this study. All five antibodies showed expected positive signal in glandular epithelium. (A) Relxin-H1 (FL-185), (B) Relxin-H1 (RLX H1), (C) Relxin H1/2 (N-18), (D) Relxin H2 (RLX H2), (E) RXFP1.

---

## Appendix B. Supplementary material paper III

---

## Supplementary Material

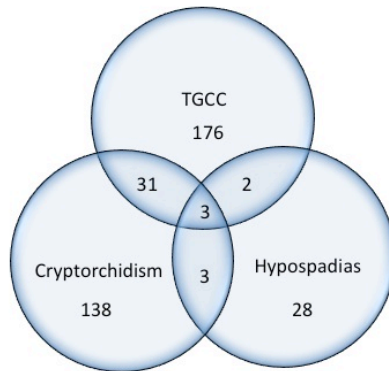
### Study cohort

The discovery cohort was of Danish ancestry and included TDS cases and military conscripts. All patients participated in a survey on semen quality among Danish men during the years from 1995 to 2009. Patients were recruited at the department of Growth and Reproduction at Rigshospitalet, providing informed consent. The group of infertile men had a mean age of 34 (21-49) and an average sperm concentration of 3.51 (0-14) million/mL, TGCC cases had a mean age of 31 (16-51). Cryptorchidism was either based on self-reporting or patient files. All hypospadias patients received surgery. The 439 controls had a mean age of 19 (17-28) and an average sperm concentration of 103 (50-451).

For the replication cohort, samples were obtained from Nordic countries and consisted of 436 cases and 235 controls, including 86 cases with TGCC, 69 with cryptorchidism and 184 controls from Denmark had a mean age of 19 (18-26) and an average sperm concentration of 112 (61-392). The 247 TGCC patients from Sweden were collected in Malmö and Stockholm (112 and 135, respectively) with an average age of 33 years (18-51). Cases with cryptorchidism (n=34) and controls (n=51) from Finland were recruited in the Turku area, all examined for cryptorchidism at birth.

**Supplementary Table 1.** Tumor type distribution among TGCC cases of the discovery cohort.

	Discovery cohort	Replication cohort
<b>Controls</b>	439	235
<b>Infertile</b>	107	0
<b>TGCC</b>	212	333
<b>Tumor type</b>		
- Seminoma	107	128
- Non-seminoma	85	153
- CIS	13	4
- No information	7	48
<b>Hypospadias</b>	31	0
<b>Cryptorchidism</b>	138	103



**Supplementary Figure 1.** Venn diagram showing the overlap of TDS phenotypes in discovery cohort: 31 patients were affected by both cryptorchidism and TGCC; 3 patients showed a shared incidence of cryptorchidism and hypospadias; and 2 patients were diagnosed with TGCC and hypospadias. 3 patients were affected by all three phenotypes.



**Supplementary Table 2.** Overview of all markers selected for replication from various approaches. We tabulate P-values (P), odds ratios (OR), and corresponding confidence intervals (OR CI) for the genome-wide association study on the testicular dysgenesis syndrome as a whole (TDS), and for all sub-phenotypes, testicular germ cell cancer (TGCC), cryptorchidism (Crypt.), infertile patients (Infer.), and hypospadias (Hypo.) separately. In addition we show the risk allele (RA) and the risk allele frequency (RAF), and which optimal genetic model (OMG) was used to test for association.

Gene	Marker	Phase	OG M	RA	RAF	adj. P	TDS P	TDS OR	TDS OR CI	TGCC P	TGCC OR	TGCC OR CI	Crypt. P	Cryp to OR	Crypto. OR CI	Infer. P	Infer. OR	Infer. OR CI	Hypo. P	Hyp o. OR	Hypo. OR CI
<i>BMP4</i>	rs17126540	Discovery	dom	G	0.082	1.0E+00	3.3E-03	1.69	(1.19-2.39)	9.7E-03	1.76	(1.14-2.71)	6.5E-03	1.93	(1.19-3.08)	3.4E-01	1.31	(0.74-2.23)	3.4E-01	1.59	(0.56-3.90)
<i>BMP4</i>	rs17126540	Replication	dom	G	0.126	1.0E+00	7.3E-01	0.93	(0.63-1.39)	9.9E-01	1.00	(0.66-1.52)	3.6E-01	0.75	(0.39-1.37)	-	-	-	-	-	-
<i>BMP6</i>	rs6913143	Discovery	rec	T	0.880	1.0E+00	5.4E-04	1.87	(1.31-2.68)	4.6E-03	2.05	(1.27-3.44)	9.9E-02	1.55	(0.94-2.68)	5.3E-02	1.80	(1.02-3.37)	8.9E-02	3.56	(1.03-22.42)
<i>BMP6</i>	rs6913143	Replication	rec	T	0.892	1.0E+00	8.1E-01	1.05	(0.69-1.59)	8.9E-01	1.03	(0.66-1.60)	7.3E-01	1.12	(0.61-2.15)	-	-	-	-	-	-
<i>BMP7</i>	rs388286	Discovery	add	C	0.468	1.0E+00	2.3E-03	1.36	(1.11-1.67)	1.6E-02	1.38	(1.06-1.79)	7.9E-03	1.49	(1.11-2.01)	2.3E-01	1.21	(0.89-1.66)	4.7E-01	1.24	(0.70-2.22)
<i>BMP7</i>	rs388286	Replication	add	C	0.472	9.9E-01	4.1E-02	1.28	(1.01-1.62)	1.7E-02	1.37	(1.06-1.78)	6.3E-01	1.09	(0.78-1.52)	-	-	-	-	-	-
<i>BMPER</i>	rs318576	Discovery	dom	G	0.656	1.0E+00	2.0E-04	2.38	(1.52-3.80)	7.8E-03	2.34	(1.29-4.58)	9.8E-02	1.74	(0.93-3.50)	7.5E-03	3.60	(1.54-10.52)	1.5E-01	4.46	(0.92-80.36)
<i>BMPER</i>	rs318576	Replication	dom	G	0.700	1.0E+00	1.6E-01	0.68	(0.38-1.15)	2.1E-01	0.69	(0.39-1.22)	2.2E-01	0.62	(0.30-1.36)	-	-	-	-	-	-
<i>BMPR1B</i>	rs17345417	Discovery	add	A	0.885	1.0E+00	4.4E-03	1.66	(1.18-2.35)	7.0E-02	1.52	(0.98-2.41)	6.9E-02	1.62	(0.98-2.79)	5.4E-02	1.78	(1.02-3.31)	9.3E-02	3.47	(1.02-21.70)
<i>BMPR1B</i>	rs17345417	Replication	add	A	0.892	1.0E+00	8.9E-01	1.03	(0.70-1.50)	6.0E-01	1.12	(0.73-1.71)	5.0E-01	0.83	(1.44-0.49)	-	-	-	-	-	-
<i>CCDC59</i>	rs11115212	Discovery	rec	C	0.225	4.2E-01	1.5E-06	5.26	(2.58-12.18)	6.8E-04	4.51	(1.94-11.31)	1.1E-05	6.97	(3.02-17.48)	9.0E-03	3.82	(1.37-10.62)	1.0E-03	8.50	(2.14-29.29)
<i>CCDC59</i>	rs11115212	Replication	rec	C	0.249	1.0E+00	2.1E-01	0.67	(0.35-1.27)	2.6E-01	0.68	(0.34-1.33)	3.9E-01	0.64	(0.21-1.64)	-	-	-	-	-	-
<i>CCDC59</i>	rs11115213	Discovery	rec	T	0.222	4.5E-01	1.1E-06	5.26	(2.58-12.20)	1.1E-03	4.29	(1.83-10.84)	2.1E-05	6.67	(2.86-16.82)	2.6E-03	4.50	(1.68-12.28)	9.2E-04	8.64	(2.17-29.75)
<i>CCDC59</i>	rs11115213	Replication	rec	T	0.258	1.0E+00	1.5E-01	0.63	(0.33-1.18)	1.8E-01	0.63	(0.32-1.23)	3.6E-01	0.62	(0.20-1.59)	-	-	-	-	-	-
<i>EPHA3</i>	rs11720651	Discovery	add	A	0.150	1.0E+00	4.3E-03	1.46	(1.12-1.90)	5.4E-02	1.40	(0.99-1.98)	1.4E-01	1.34	(0.90-1.96)	1.7E-03	1.87	(1.26-2.75)	6.3E-01	1.22	(0.52-2.64)
<i>EPHA3</i>	rs11720651	Replication	add	A	0.147	1.0E+00	2.2E-01	1.24	(0.88-1.75)	3.1E-01	1.21	(0.84-1.76)	2.7E-01	1.31	(0.80-2.11)	-	-	-	-	-	-
<i>EPHB2</i>	rs12723359	Discovery	dom	A	0.674	1.0E+00	5.7E-05	2.71	(1.61-4.72)	3.3E-02	2.10	(1.10-4.37)	2.2E-02	2.79	(1.25-7.42)	1.2E-02	4.59	(1.64-19.18)	2.4E-01	3.34	(0.68-60.41)
<i>EPHB2</i>	rs12723359	Replication	dom	A	0.680	1.0E+00	2.3E-01	0.72	(0.41-1.23)	4.1E-01	0.78	(0.43-1.40)	1.4E-01	0.57	(0.28-1.22)	-	-	-	-	-	-
<i>FOXP1</i>	rs2180892	Discovery	dom	G	0.740	1.0E+00	7.1E-05	4.35	(2.14-9.80)	9.7E-03	4.02	(1.57-13.66)	2.9E-02	3.81	(1.34-16.06)	3.7E-02	4.65	(1.38-28.98)	9.9E-01	NA	(0.00-NA)
<i>FOXP1</i>	rs2180892	Replication	dom	G	0.770	1.0E+00	7.1E-01	0.87	(0.42-1.74)	7.7E-01	0.90	(0.42-1.88)	6.8E-01	0.81	(0.31-2.36)	-	-	-	-	-	-
<i>GIAS/NBPF14</i>	rs1495956	Discovery	dom	A	0.228	1.0E+00	8.8E-05	1.74	(1.32-2.29)	7.7E-03	1.62	(1.14-2.31)	8.5E-03	1.71	(1.15-2.55)	3.2E-04	2.23	(1.45-3.47)	7.2E-01	1.16	(0.51-2.58)
<i>GIAS/NBPF14</i>	rs1495956	Replication	dom	A	0.253	1.0E+00	2.8E-01	0.83	(0.60-1.16)	3.9E-01	0.86	(0.60-1.22)	2.9E-01	0.76	(0.46-1.25)	-	-	-	-	-	-
<i>HOXD13</i>	rs1522830	Discovery	dom	C	0.086	1.8E-01	4.9E-07	2.34	(1.68-3.29)	3.1E-06	2.64	(1.75-3.97)	5.8E-05	2.54	(1.61-3.99)	4.8E-02	1.69	(0.99-2.81)	6.1E-02	2.31	(0.91-5.38)
<i>HOXD13</i>	rs1522830	Replication	dom	C	0.116	1.0E+00	8.3E-01	0.96	(0.63-1.45)	8.1E-01	0.95	(0.61-1.48)	9.5E-01	0.98	(0.52-1.78)	-	-	-	-	-	-
<i>HOXD13</i>	rs16863266	Discovery	dom	A	0.086	2.4E-01	5.8E-07	2.33	(1.67-3.27)	4.8E-06	2.60	(1.72-3.91)	5.8E-05	2.54	(1.61-3.99)	4.8E-02	1.69	(0.99-2.81)	6.1E-02	2.31	(0.91-5.38)
<i>HOXD13</i>	rs16863266	Replication	dom	A	0.121	1.0E+00	4.4E-01	0.85	(0.56-1.29)	3.9E-01	0.82	(0.52-1.29)	7.8E-01	0.92	(0.49-1.66)	-	-	-	-	-	-
<i>HOXD13</i>	rs17198418	Discovery	dom	G	0.077	2.7E-01	6.3E-07	2.36	(1.67-3.38)	2.0E-05	2.53	(1.65-3.88)	3.8E-05	2.68	(1.67-4.27)	5.9E-02	1.69	(0.97-2.88)	2.7E-02	2.71	(1.07-6.35)
<i>HOXD13</i>	rs17198418	Replication	dom	G	0.105	1.0E+00	7.0E-01	0.92	(0.60-1.41)	6.8E-01	0.91	(0.58-1.43)	8.9E-01	0.95	(0.50-1.77)	-	-	-	-	-	-
<i>HOXD13</i>	rs17198432	Discovery	dom	A	0.073	2.2E-02	4.7E-08	2.58	(1.82-3.70)	1.4E-06	2.85	(1.86-4.38)	8.9E-06	2.90	(1.81-4.63)	4.2E-02	1.76	(1.01-3.01)	2.1E-02	2.83	(1.11-6.64)
<i>HOXD13</i>	rs17198432	Replication	dom	A	0.107	1.0E+00	8.8E-01	0.97	(0.64-1.48)	8.5E-01	0.96	(0.61-1.50)	1.0E+00	1.00	(0.52-1.83)	-	-	-	-	-	-
<i>HOXD13</i>	rs2437901	Discovery	dom	G	0.096	9.1E-01	5.2E-06	2.14	(1.54-2.98)	3.3E-05	2.35	(1.57-3.52)	1.3E-04	2.39	(1.53-3.74)	1.0E-01	1.54	(0.91-2.55)	9.5E-02	2.10	(0.83-4.89)
<i>HOXD13</i>	rs2437901	Replication	dom	G	0.122	1.0E+00	6.2E-01	0.91	(0.61-1.35)	7.6E-01	0.94	(0.61-1.43)	5.1E-01	0.81	(0.43-1.47)	-	-	-	-	-	-

HOXD13	rs7574075	Discovery	dom	G	0.085	2.2E-01	5.8E-07	2.34	(1.67-3.29)	4.0E-06	2.62	(1.74-3.96)	4.2E-05	2.59	(1.64-4.07)	6.7E-02	1.63	(0.95-2.73)	5.6E-02	2.35	(0.93-5.49)
HOXD13	rs7574075	Replication	dom	G	0.109	1.0E+00	7.1E-01	0.92	(0.61-1.41)	6.2E-01	0.89	(0.57-1.40)	9.4E-01	1.02	(0.54-1.86)	-	-	-	-	-	-
HOXD13	rs7586946	Discovery	dom	G	0.077	5.4E-01	1.9E-06	2.34	(1.65-3.35)	2.0E-05	2.53	(1.65-3.88)	3.8E-05	2.68	(1.67-4.27)	9.6E-02	1.60	(0.91-2.73)	2.7E-02	2.71	(1.07-6.35)
HOXD13	rs7586946	Replication	dom	G	0.106	1.0E+00	9.6E-01	1.01	(0.66-1.55)	8.5E-01	1.04	(0.66-1.65)	7.9E-01	0.92	(0.48-1.70)	-	-	-	-	-	-
ITPR1	rs2259801	Discovery	rec	A	0.590	1.0E+00	8.6E-05	1.71	(1.30-2.27)	1.7E-03	1.77	(1.24-2.54)	2.6E-02	1.58	(1.05-2.38)	1.5E-02	1.72	(1.11-2.65)	9.3E-02	2.01	(0.89-4.58)
ITPR1	rs2259801	Replication	rec	A	0.615	1.0E+00	1.1E-01	1.32	(0.94-1.87)	9.2E-02	1.37	(0.95-1.99)	4.7E-01	1.20	(0.73-1.98)	-	-	-	-	-	-
KITLG	rs11104952	Discovery	add	G	0.788	1.0E+00	3.3E-03	1.48	(1.16-1.90)	1.2E-05	2.33	(1.62-3.45)	4.9E-01	1.13	(0.81-1.59)	2.7E-01	1.23	(0.86-1.81)	7.5E-01	1.12	(0.59-2.34)
KITLG	rs11104952	Replication	add	G	0.807	1.4E-03	6.0E-05	1.95	(1.41-2.71)	5.7E-06	2.39	(1.65-3.52)	3.5E-01	1.24	(0.80-1.97)	-	-	-	-	-	-
KITLG	rs1352947	Discovery	add	T	0.808	1.0E+00	3.1E-03	1.52	(1.18-1.97)	3.1E-06	2.74	(1.82-4.26)	6.5E-01	1.08	(0.77-1.55)	2.5E-01	1.26	(0.86-1.90)	1.0E+00	1.00	(0.52-2.11)
KITLG	rs1352947	Replication	add	T	0.819	2.0E-03	8.5E-05	1.93	(1.39-2.69)	2.6E-05	2.23	(1.55-3.28)	1.7E-01	1.38	(0.89-2.24)	-	-	-	-	-	-
KITLG	rs3782181	Discovery	add	A	0.788	1.0E+00	2.2E-03	1.49	(1.17-1.90)	1.1E-05	2.35	(1.63-3.49)	4.9E-01	1.13	(0.81-1.60)	2.7E-01	1.24	(0.86-1.82)	7.5E-01	1.12	(0.59-2.35)
KITLG	rs3782181	Replication	add	A	0.801	1.2E-03	4.9E-05	1.94	(1.41-2.69)	4.0E-06	2.40	(1.67-3.52)	3.5E-01	1.23	(0.81-1.93)	-	-	-	-	-	-
KITLG	rs4474514	Discovery	add	A	0.790	1.0E+00	4.2E-03	1.48	(1.16-1.89)	1.3E-05	2.33	(1.61-3.46)	5.2E-01	1.12	(0.80-1.58)	2.8E-01	1.23	(0.85-1.80)	7.7E-01	1.11	(0.58-2.33)
KITLG	rs4474514	Replication	add	A	0.805	6.8E-03	2.8E-04	1.81	(1.32-2.50)	4.3E-05	2.14	(1.49-3.10)	3.6E-01	1.23	(0.80-1.94)	-	-	-	-	-	-
MEIS1	rs12470855	Discovery	dom	C	0.706	1.0E+00	1.5E-03	2.48	(1.44-4.40)	9.2E-02	1.81	(0.94-3.79)	8.3E-03	4.98	(1.77-20.82)	1.4E-01	1.94	(0.86-5.22)	9.9E-01	NA	(0.00-NA)
MEIS1	rs12470855	Replication	dom	C	0.716	1.0E+00	4.7E-01	1.26	(0.66-2.34)	9.3E-01	1.03	(0.54-1.95)	8.0E-02	3.75	(1.06-23.84)	-	-	-	-	-	-
MIPEP	rs4769283	Discovery	dom	G	0.641	1.0E+00	3.5E-04	2.41	(1.55-3.83)	8.8E-03	2.26	(1.26-4.31)	1.3E-02	2.52	(1.27-5.58)	3.6E-02	2.28	(1.11-5.30)	1.4E-01	4.65	(0.96-83.82)
MIPEP	rs4769283	Replication	dom	G	0.681	1.0E+00	7.3E-01	0.91	(0.51-1.58)	6.3E-01	0.87	(0.47-1.56)	9.1E-01	1.05	(0.46-2.59)	-	-	-	-	-	-
NRP2	rs849515	Discovery	add	G	0.530	1.0E+00	3.9E-04	1.42	(1.16-1.73)	2.8E-02	1.33	(1.03-1.73)	4.0E-04	1.71	(1.28-2.31)	2.3E-01	1.21	(0.89-1.64)	4.6E-02	1.85	(1.03-3.47)
NRP2	rs849515	Replication	add	G	0.569	1.0E+00	4.7E-01	1.09	(0.86-1.39)	1.9E-01	1.19	(0.92-1.54)	4.4E-01	0.87	(1.23-0.62)	-	-	-	-	-	-
PLXNC1	rs2291333	Discovery	dom	C	0.086	7.3E-01	2.9E-06	2.24	(1.60-3.16)	2.7E-06	2.67	(1.77-4.04)	5.1E-03	1.97	(1.22-3.15)	1.1E-02	1.94	(1.15-3.21)	1.4E-01	1.99	(0.75-4.74)
PLXNC1	rs2291333	Replication	dom	C	0.118	1.0E+00	7.9E-01	0.95	(0.63-1.43)	7.3E-01	0.93	(0.60-1.43)	9.7E-01	1.01	(0.55-1.82)	-	-	-	-	-	-
PTGFR	rs12742293	Discovery	rec	T	0.258	1.0E+00	2.5E-05	3.14	(1.81-5.77)	3.7E-04	3.34	(1.73-6.60)	1.5E-02	2.58	(1.18-5.50)	3.6E-04	3.85	(1.82-8.12)	3.9E-01	1.95	(0.30-7.40)
PTGFR	rs12742293	Replication	rec	T	0.286	1.0E+00	4.4E-01	1.26	(0.71-2.29)	2.7E-01	1.41	(0.77-2.61)	6.8E-01	0.82	(0.29-2.01)	-	-	-	-	-	-
RTNMR	rs1640351	Discovery	rec	C	0.142	1.0E+00	4.2E-04	7.89	(2.24-49.99)	2.7E-03	10.64	(2.71-70.31)	1.4E-02	7.90	(1.68-55.65)	6.3E-02	5.53	(0.90-42.40)	9.9E-01	0.00	(NA-NA)
RTNMR	rs1640351	Replication	rec	C	0.120	1.0E+00	1.0E+00	1.00	(0.28-3.94)	8.9E-01	1.10	(0.29-4.49)	7.3E-01	0.68	(0.03-4.68)	-	-	-	-	-	-
SEMA3A	rs7808864	Discovery	rec	G	0.744	1.0E+00	1.7E-04	1.77	(1.33-2.34)	1.4E-02	1.57	(1.10-2.27)	6.8E-03	1.78	(1.18-2.72)	1.5E-03	2.13	(1.35-3.43)	1.5E-01	1.89	(0.83-4.70)
SEMA3A	rs7808864	Replication	rec	G	0.781	1.0E+00	5.8E-01	0.91	(0.65-1.27)	4.9E-01	0.88	(0.62-1.26)	9.8E-01	1.01	(0.61-1.67)	-	-	-	-	-	-
SEMA3C	rs11768393	Discovery	add	G	0.928	1.0E+00	3.6E-03	1.41	(0.96-2.09)	6.2E-02	1.69	(1.00-3.05)	9.0E-01	1.03	(0.63-1.77)	3.4E-01	1.35	(0.76-2.66)	2.2E-01	3.31	(0.80-56.69)
SEMA3C	rs11768393	Replication	add	A	0.056	1.0E+00	5.4E-01	1.16	(0.73-1.90)	9.7E-01	1.01	(0.60-1.70)	1.2E-01	1.65	(0.86-3.07)	-	-	-	-	-	-
TGFB2	rs12042727	Discovery	dom	G	0.047	1.0E+00	3.3E-03	1.96	(1.29-3.04)	8.7E-04	2.36	(1.42-3.93)	2.0E-01	1.50	(0.79-2.74)	4.0E-02	1.92	(1.01-3.54)	2.9E-01	1.83	(0.51-5.11)
TGFB2	rs12042727	Replication	dom	G	0.082	1.0E+00	1.3E-01	0.69	(0.43-1.12)	1.5E-01	0.68	(0.41-1.14)	3.8E-01	0.73	(0.34-1.45)	-	-	-	-	-	-
TGFBR3	rs12082710	Discovery	rec	T	0.580	1.0E+00	2.4E-04	1.77	(1.33-2.36)	2.0E-02	1.55	(1.07-2.23)	2.2E-02	1.62	(1.07-2.44)	1.0E-04	2.38	(1.54-3.70)	8.1E-02	2.04	(0.90-4.56)
TGFBR3	rs12082710	Replication	rec	T	0.587	3.8E-01	1.6E-02	1.52	(1.08-2.15)	3.1E-02	1.49	(1.04-2.14)	5.4E-02	1.64	(0.99-2.71)	-	-	-	-	-	-
WNT5A	rs1380119	Discovery	rec	A	0.247	1.0E+00	1.2E-03	2.72	(1.55-5.03)	1.5E-03	2.98	(1.52-5.94)	1.4E-03	3.25	(1.56-6.74)	7.7E-02	2.15	(0.88-4.91)	9.4E-01	0.93	(0.05-4.83)
WNT5A	rs1380119	Replication	rec	A	0.180	1.0E+00	3.8E-01	1.47	(0.65-3.63)	2.8E-01	1.62	(0.69-4.10)	9.9E-01	0.99	(0.21-3.52)	-	-	-	-	-	-

**Supplementary Table 3.** Integration of complementary data types used to select markers for validation. The table shows 14 genes that were selected based on the ISB: Integrative Systems Biology, Pathway: pathway analysis, MGI.PPI: layer based on data from the Mouse Genome Informatics database (MGI) and protein-protein interaction data, Devel.expr: layer based on developmental expression data, GWAS: layer based on univariate association from genome-wide association study.

Gene	Marker	Combined data types	Combined rank	Selection criteria
<i>TGFBR3</i>	rs12082710	GWAS, MGI.PPI, Devel.expr	3	ISB
<i>ID2</i>	rs2630720	GWAS, MGI.PPI, Devel.expr	4	ISB
<i>ITPR1</i>	rs2259801	GWAS, MGI.PPI, Devel.expr	5	ISB
<i>MEIS1</i>	rs12470855	GWAS, MGI.PPI, Devel.expr	6	ISB
<i>WNT5A<sup>a</sup></i>	rs1380119	GWAS, MGI.PPI, Devel.expr,*	7	ISB
<i>EPHB2</i>	rs12723359	GWAS, MGI.PPI, Devel.expr	11	ISB
<i>BMP6</i>	rs6913143	GWAS, MGI.PPI	3	Pathway
<i>BMP7</i>	rs388286	GWAS, MGI.PPI, Devel.expr.human	3	Pathway
<i>NRP2</i>	rs849515	GWAS, Devel.expr.human	6	Pathway
<i>BMP4</i>	rs17126540	GWAS, MGI.PPI, Devel.expr	29	Pathway
<i>FOXG1</i>	rs2180892	GWAS, Devel.expr.human	1	Shared interaction partner with ID2
<i>TGFB2</i>	rs12042727	GWAS, MGI.PPI, Devel.expr.human	14	Ligand to TGFBR3
<i>BMPR1B</i>	rs17345417	GWAS, MGI.PPI, Devel.expr.human	17	Receptor of BMPs
<i>EPHA3</i>	rs11720651	GWAS, MGI.PPI, Devel.expr.human	129	Share ligand with EPHB2 (EFNB2)

<sup>a</sup> *WNT5A* was identified using a fourth data type in the integrative approach. The later was based on microarray expression data of microdissected testicular carcinoma in situ cells (CIS) and whole adult testis were used to identify genes differentially expressed in testicular CIS genes by a moderated t-test (empirical Bayes).

**Supplementary Table 4.** Ranking of genes selected for validation according to different layers used in the study. The layers used to rank all human genes include a layer based on the genome-wide association data, a layer of genes prioritized using protein-protein interaction and mouse knock-out data, a layer based on meta-analysis of gene expression data from four independent studies: Small, McMahon, Houmard, and Sonne, an a layer based on expression in carcinoma in situ cells (CIS). P: P-value, T: targeted knock-out, P (Bonf): P-value after Bonferroni multiple testing correction, P(BH): P-values after Benjamini-Hochberg correction for multiple testing to control false discovery rate.

	GWAS (SNP)							MGI.PPI <sup>11</sup>			Small <sup>22</sup>		McMahon <sup>33</sup>		Houmard <sup>44</sup>		Sonne <sup>55</sup>			
	Marker	Position	Number Eff. tests	P Sidak	Rank Sidak	P-value	Rank	T	P(Bonf)	Rank	P(BH)	Rank	P(BH)	Rank	CV	Rank	Anova P (BH)	Anova rank	P (BH)	Rank
TGFBR3	rs12082710	91927925	28.9728	7.07E-03	238	2.45E-04	111	0	5.34E-02	42	8.97E-03	160	1.14E-03	4132	4.70E-02	1701	4.88E-01	14954	7.97E-01	8409
ID2	rs2630720	8488928	53.6116	1.63E-02	569	3.06E-04	130	0	1.08E+01	218	1.22E-02	283	9.54E-04	3930	5.93E-02	764	3.87E-02	422	8.93E-01	13391
ITPR1	rs2259801	4488543	48.3051	4.16E-03	141	8.63E-05	50	0	9.76E+02	1284	1.84E-02	577	4.39E-05	1248	4.71E-02	1682	3.29E-01	11002	2.67E-01	511
MEIS1	rs12470855	66425677	38.1968	5.53E-02	1751	1.49E-03	601	0	2.63E+02	654	7.88E-03	131	6.80E-05	1512	6.22E-02	641	5.45E-02	798	2.53E-01	432
WNT5A	rs1380119	55128254	40.8407	4.70E-02	1521	1.18E-03	485	1	1.14E+03	1455	7.24E-02	3354	1.88E-04	2272	2.25E-02	11625	2.02E-02	139	2.03E-01	264
EPHB2	rs12723359	23010348	30.6653	1.74E-03	58	5.67E-05	30	1	5.39E+01	355	6.55E-02	3021	4.62E-04	3162	4.05E-02	2769	1.14E-01	2915	6.93E-01	4694
BMP6	rs6913143	7769235	32.322	1.74E-02	610	5.44E-04	251	0	1.92E-01	64	2.59E-02	985	1.27E-03	4260	1.64E-02	15853	5.19E-01	15541	7.24E-01	5332
BMP7	rs388286	54898831	54.5674	1.16E-01	3285	2.26E-03	889	1	8.99E-01	105	5.97E-02	2780	2.42E-03	5051	1.05E-01	77	4.53E-02	596	4.41E-01	1447
NRP2	rs849515	206220818	34.4264	1.33E-02	459	3.88E-04	166	0	NA	NA	1.68E-01	7121	4.91E-05	1304	4.63E-02	1797	2.81E-02	240	8.43E-01	10992
BMP4	rs17126540	52994319	39.4787	1.22E-01	3414	3.28E-03	1230	1	6.87E-01	97	2.88E-02	1116	5.09E-03	6053	2.41E-02	10442	1.27E-01	3440	9.25E-01	14827
FOXC1	rs2180892	28000063	24.2286	1.72E-03	56	7.10E-05	37	0	8.82E+02	1241	2.73E-01	9955	4.23E-01	13435	9.73E-02	107	7.54E-02	1491	5.37E-01	2427
TGFB2	rs12042727	216736203	15.1578	4.86E-02	1559	3.28E-03	1231	1	3.88E+02	779	1.58E-01	6823	4.86E-01	13690	4.39E-02	2151	3.34E-02	285	9.17E-01	14415
BMPRI1B	rs17345417	96167509	24.6353	1.02E-01	2966	4.36E-03	1541	0	9.47E-02	51	1.32E-01	5912	1.52E-03	4452	6.22E-02	636	1.13E-01	2845	8.96E-01	13571
EPHA3	rs11720651	89343751	26.1186	1.06E-01	3051	4.27E-03	1522	0	5.61E+03	3357	9.03E-03	177	2.26E-02	8249	5.72E-02	880	5.48E-02	826	8.70E-01	12342

1. Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. & Blake, J.A. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* **36**, D724-728 (2008).
2. Small, C.L., Shima, J.E., Uzumcu, M., Skinner, M.K. & Griswold, M.D. Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol. Reprod* **72**, 492-501 (2005).
3. McMahon, A.P. et al. GUDMAP: the genitourinary developmental molecular anatomy project. *J. Am. Soc. Nephrol* **19**, 667-671 (2008).
4. Houmard, B. et al. Global gene expression in the human fetal testis and ovary. *Biol. Reprod* **81**, 438-443 (2009).
5. Sonne, S.B. et al. Analysis of gene expression profiles of microdissected cell populations indicates that testicular carcinoma in situ is an arrested gonocyte. *Cancer Res* **69**, 5241-5250 (2009).

**Supplementary Table 5.** Mammalian Phenotype terms were used to select genes that are relevant for TDS from the Mouse Genomics Informatics database. The table shows a set of 34 TDS related morphologies, which were curated manually according to developmental defects of the testis and relation to TDS. Since spermatogenesis is a complex process involving a plethora of genes, which is not directly related to TDS, we excluded all genes in the 'abnormal spermatogenesis' branch of the MGI ontology. Another term, 'male infertility' includes many genes that are not specifically related to TDS, and was excluded as well. All descendant phenotype terms of this set of curated terms were then retrieved from the Mammalian Phenotype ontology, resulting in a total of 122 terms.

Phenotype ID	Phenotype term
MP:0003826	abnormal Mullerian duct morphology
MP:0003827	abnormal Wolffian duct morphology
MP:0004728	abnormal efferent ductules of testis
MP:0002631	abnormal epididymis morphology
MP:0009199	abnormal external male genitalia morphology
MP:0005651	abnormal gonad rudiment morphology
MP:0005149	abnormal gubernaculum morphology
MP:0008016	abnormal male inguinal canal morphology
MP:0003315	abnormal perineum morphology
MP:0002982	abnormal primordial germ cell migration
MP:0008391	abnormal primordial germ cell morphology
MP:0008390	abnormal primordial germ cell proliferation
MP:0006416	abnormal rete testis morphology
MP:0002216	abnormal seminiferous tubule morphology
MP:0003125	abnormal septation of the cloaca
MP:0002685	abnormal spermatogonia proliferation
MP:0003830	abnormal testis development
MP:0002769	abnormal vas deferens morphology
MP:0009206	absent internal male genitalia
MP:0006415	absent testes
MP:0002286	cryptorchism
MP:0009207	internal male genitalia hypoplasia
MP:0002789	male pseudohermaphroditism
MP:0002996	Ovotestis
MP:0003129	persistent cloaca
MP:0002995	primary sex reversal
MP:0001939	secondary sex reversal
MP:0005652	sex reversal
MP:0002214	streak gonad
MP:0001940	testicular hypoplasia
MP:0002213	true hermaphroditism

## Complex ranking

The rationale behind analysis of a set of proteins that exerts its function as a protein complex is that genetic variations may perturb different components of the complex. While a single perturbation of one gene does not necessarily affect the protein complex as a whole, a concerted perturbation of several components is much more likely to interfere with proper functioning of the protein complex, and potentially disrupts its biological role within the cell.

We compiled a list of all human protein complexes using protein-protein interaction data, and tested each complex for the joint effect of association signals from SNPs located at the genes of a complex. Ranked fourth, was a set of interacting proteins that are members of the TGF- $\beta$  superfamily; BMPER, BMP2, BMP7, BMP6 and BMP4.

**Supplementary Table 6.** Complexes and P-values for the most significant complexes from the analysis described above. The table shows the four most significant complexes, consisting of 4, 3, 3, and 5 proteins, respectively.

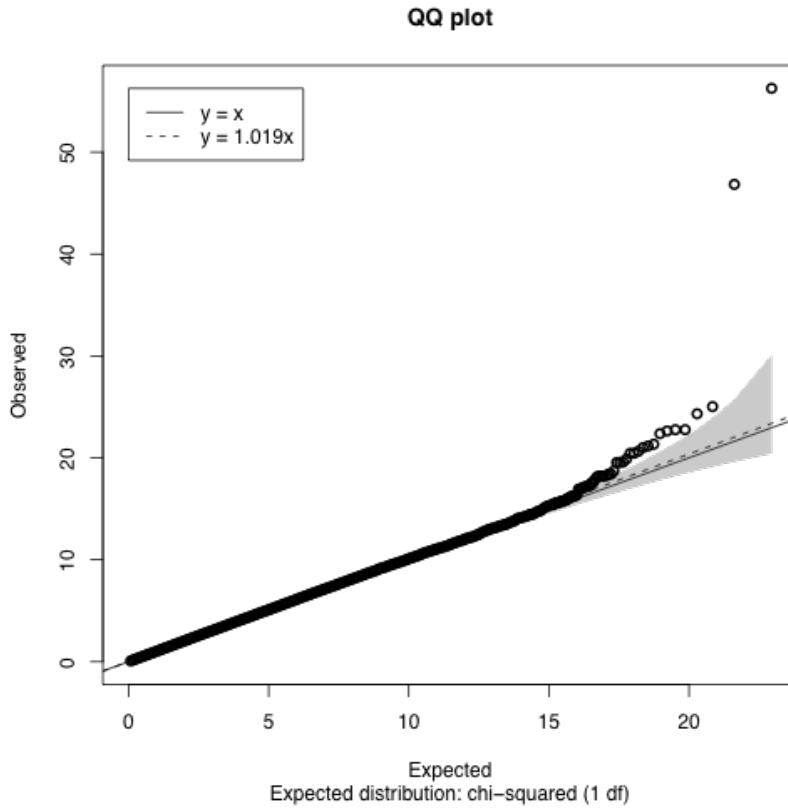
Rank	Gene	SNP ID	p-value
1	<i>SEMA3B</i>	rs6776145	0.038540
	<i>NPR2</i>	rs849515	0.000388
	<i>NPR1</i>	rs10827285	0.000020
	<i>SEMA3A</i>	rs7808864	0.000174
2	<i>ITPR1</i>	rs2259801	0.000086
	<i>RTN4R</i>	rs1640351	0.000420
	<i>TNFRSF19</i>	rs4769283	0.000347
3	<i>SEMA3C</i>	rs11768393	0.003561
	<i>NPR2</i>	rs849515	0.000388
	<i>NPR1</i>	rs10827285	0.000020
4	<i>BMP2</i>	rs6038644	0.000993
	<i>BMP4</i>	rs17126540	0.003277
	<i>BMP6</i>	rs6913143	0.000544
	<i>BMP7</i>	rs388286	0.002264
	<i>BMPER</i>	rs318576	0.000198

**Supplementary Table 7.** Genotype proportions rs17198432, the marker at the HOXD gene for cases, controls and all TDS sub-phenotypes separately.

Genotype	Controls	Cases	Cryptorchidism	Hypospadias	Infertile	Testiscancer
AA	0.01	0.02	0.02	0.04	0.01	0.02
AB	0.13	0.26	0.28	0.28	0.19	0.29
BB	0.86	0.72	0.69	0.68	0.80	0.69

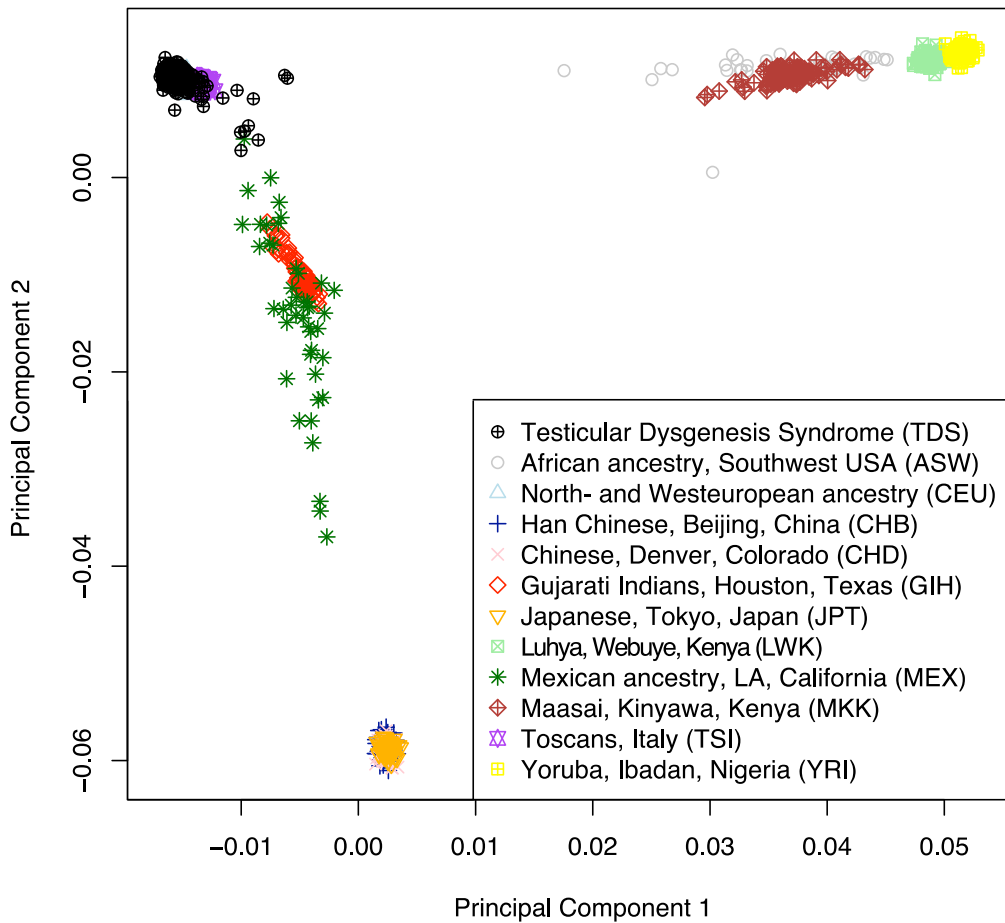
**Supplementary Table 8.** Comparison of allele frequencies for rs17198432, which was found in the HOXD/LNP region. The table shows that the minor allele frequency varies between different ethnic groups.

<b>Population</b>	<b>A allele</b>	<b>C allele</b>
Gujarati Indians in Houston. Texas (GIH)	0.847	0.153
Utah residents with Northern and Western European ancestry (CEU)	0.872	0.128
Danish population used in this study. Denmark	0.883	0.117
Toscans in Italy (TSI)	0.886	0.114
Mexican ancestry in Los Angeles. California (MEX)	0.890	0.110
Maasai in Kinyawa. Kenya (MKK)	0.972	0.028
Luhya in Webuye. Kenya (LWK)	0.989	0.011
African ancestry in Southwest USA (ASW)	0.981	0.019
Han Chinese in Beijing. China (CHB)	1.000	0.000
Yoruba in Ibadan. Nigeria (YRI)	1.000	0.000
Japanese in Tokyo. Japan (JPT)	1.000	0.000
Chinese in Metropolitan Denver. Colorado (CHD)	NA	NA



**Supplementary Figure 2.** Quantile-quantile plot showing expected versus observed test statistics. The genomic inflation factor of 1.019 indicates very little confounding.





**Supplementary Figure 3.** Principal Components Analysis (PCA) of 11 HapMap phase III populations and the Danish discovery cohort used in this study.

---

## Appendix C. Supplementary material paper IV

---

**Supplementary Table 1**

Gene sets with an association of rare CNVs.

<i>PTPN1</i>	<i>present</i>	Gene set type	Term name	Posterior	Local	Cases	Controls	OR	OR.CI	N	N	Term ID
	NO			FDR	FDR	%	%			cases	controls	
		GO Biological Process	Regulation of cell migration	0.98	0.021	1.8%	1.1%	3.47	(1.12-11.82)	10	6	go:0030334
		Protein complex	NTRK1	0.98	0.023	2.6%	1.4%	4.00	(1.56-11.10)	15	8	ensg000000198400
		NCI-Nature Curated Pathway	Calcineurin-regulated NFAT-dependent transcription in lymphocytes	0.98	0.024	1.2%	0.2%	14.51	(1.84-656.47)	7	1	ncipid:200048
		GO Molecular Function	Hydrolase activity, acting on ester bonds	0.97	0.027	5.3%	6.5%	1.75	(1.00-3.02)	30	37	go:0016788
		Protein complex	CLK2	0.97	0.028	1.4%	0.5%	5.54	(1.31-32.78)	8	3	ensg000000176444
		Protein complex	PSTPIP1	0.97	0.031	1.2%	0.5%	4.82	(1.08-29.22)	7	3	ensg000000140368
		Protein complex	NTRK2	0.97	0.031	1.1%	0.2%	12.38	(1.48-571.36)	6	1	ensg000000148053
		NCI-Nature Curated Pathway	Calcineurin-regulated NFAT-dependent transcription in lymphocytes	0.96	0.039	1.4%	0.2%	16.68	(2.21-742.43)	8	1	ncipid:200048
		Protein complex	PTPN1	0.96	0.040	1.4%	0.5%	5.54	(1.31-32.78)	8	3	ensg000000196396
	NO	GO Biological Process	macromolecular complex disassembly	0.96	0.040	1.4%	0.5%	5.54	(1.31-32.78)	8	3	go:0032984
		Protein complex	LPP	0.95	0.047	0.9%	0.2%	10.25	(1.14-487.12)	5	1	ensg000000145012
	NO	GO Biological Process	Positive regulation of catalytic activity	0.95	0.047	2.3%	1.9%	2.47	(1.00-6.23)	13	11	go:0043085
		Reactome pathway	Integrin cell surface interactions	0.95	0.047	1.2%	0.4%	7.24	(1.36-71.98)	7	2	react_13552
		GO Molecular Function	Phosphoric ester hydrolase activity	0.95	0.048	3.9%	4.2%	1.95	(1.01-3.75)	22	24	go:0042578