Computational Prediction of Gene Targets for Fetal Alcohol Spectrum Disorders

Zané Lombard

A thesis submitted to the Faculty of Health Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy

June 2008

Supervisor: Prof Michèle Ramsay

Declaration

I, Zané Lombard, declare that this thesis is my own, unaided work, unless otherwise specified in the text. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

……..…… ……

Zané Lombard Date

Education is the great engine of personal development.
It is through education that the daughter of a peasant can
become a doctor, that the son of a mineworker can
become the head of the mine, that a child of farm workers
can become president of a great nation.
It is what we make out of what we have, not what we are given,
that separates one person from another

– Nelson Mandela, 'Long walk to freedom'

Abstract

Fetal alcohol spectrum disorders (FASD) describe the range of disorders that result from in utero alcohol exposure. FASD is a serious global health problem and is observed at exceedingly high frequencies in certain South African communities. Although in utero alcohol exposure is the primary trigger, there is evidence that genetic- and other susceptibility factors contribute towards FASD development. To date, no genome-wide association or linkage studies have been performed for any of the FASD syndromes.

The main objectives of this study were to develop an innovative approach to computationally identify biologically plausible candidate genes for FASD, for a future association study, and to evaluate the appropriateness and validity of this approach. Further, an in silico analysis of known single nucleotide polymorphisms (SNPs) within the top-ranked candidate gene was performed in conjunction with de novo SNP detection, to select a subset of SNPs based on proposed functional impact on gene expression and protein function, for a prospective association study.

A computational binary filtering technique was designed that can be employed to prioritize genes in a candidate list, or could be used to rank all genes in the genome in the absence of such a list. 10174 FASD candidate genes were initially selected from the whole genome using a previously described method. Hereafter the candidates were prioritized using a binary filtering technique. The biological enrichment of the ranked genes was assessed by investigating the protein-protein interactions, functional enrichment and common promoter element binding sites of the top-ranked genes. A group of 87 genes was prioritized as candidates highlighting many strong candidates from the TGF-β, MAPK and Hedgehog signalling pathways, which are all integral to fetal development and potential targets for alcohol's teratogenic effect.

To assess the effectiveness and accuracy of this computational approach, X-linked mental retardation (XLMR) was used as a test disease, considering that XLMR is a set of heterogeneous disorders of which some of the underlying genetics is known. This implementation resulted in a prioritized gene list with a noted enrichment of known XLMR genes among the top-ranked genes. Furthermore, the top-ranked list contained genes that were biologically relevant to XLMR, and could potentially be as yet unknown candidate genes for XLMR. Indeed, many of the top-ranked genes mapped to XLMR candidate regions, confirming their status as good candidates.

Finally, a subset of seven known and novel SNPs was selected within *FGFR1* based on putative functional impact. Data from the HapMap project was used to identify tag SNPs for *FGFR1* to complement the selection made based on function.

The main limitation of the proposed computational approach to candidate gene prediction is that it is primarily based on gene annotation, and that it is therefore biased towards selecting better-annotated genes. However, the results obtained in this study suggest that the described computational method is an effective approach that can identify likely candidates that are biologically relevant to the disease of interest, and therefore appropriate for a candidate-gene association studies. In practice, this technique is an appropriate approach to select a workable set of candidate genes for a complex disease, in a setting where a whole-genome association study is not a viable option.

.

Acknowledgements

I would like to express my gratitude to the following people and institutions:

First and foremost I want to thank my supervisor, Prof Michele Ramsay, for creating this opportunity for me to pursue my academic goals. Thank you for your guidance, support and many fruitful discussions; and for motivating me to complete this (sometimes) daunting task.

I am grateful to Prof Denis Viljoen, and the FARR team for creating opportunities for sample collection, for introducing us to the people of De Aar and Upington, and for supplying invaluable information on FASD, particularly the clinical aspects.

The National Bioinformatics Network, the Medical Research Council, the National Health Laboratory Service and the University of the Witwatersrand for personal and project financial support. I would particularly want to thank the NBN, for their support of this project, and for making my introduction to Bioinformatics a very pleasant experience.

The staff from WITS Bioinformatics and the WITS School of Computer Science, specifically Prof. Scott Hazelhurst, Dries Oelofse, Khayeni Ndlovu and Dr. Alexander Holt, for helpful technical support and teachings in computer science.

My collaborators at the South African National Bioinformatics Institute – Prof. Winston Hide, Prof. Vladimir Bajic, Dr. Nicki Tiffin, Adele Kruger and Dr. Oliver Hoffman. Your insight and unique approaches to this research problem were invaluable. I am also thankful to Dr. Janet Kelso for her input in the early stages of the project.

A special thank you to Shelley Macaulay, and Dhamari Naidoo for your friendship and efforts on the FAS project; and to Candice-lee de Carvahlo, Katpaham Shantikumar and Silke Arndt for your help and encouragement regarding the laboratory side of this project.

To my wonderful family – thank you for your constant encouragement. Thank you to my mom and dad for setting me on this path, and to my mom-in-law for her support and help with Alexia during these last months of writing up.

Thank you to my great friends for bringing balance to my life – especially Leandra, Marisha, Mnandi and Nicole.

And finally, but not least, to Nico – thank you for your never-ending love and encouragement. Your support is the single most important thing that compels me to achieve my goals. Your sheer determination to make a success of everything you take on is inspiring.

Preface

Fetal alcohol spectrum disorders (FASD) encompasses a series of serious and permanent disorders, that include anomalies such as prenatal and postnatal growth retardation, central nervous system (CNS) dysfunction, characteristic craniofacial malformation and other organ abnormalities (Clarren et al., 1978; Day et al., 1999; Sulik and Johnston 1983). Particularly in South Africa, alarmingly high prevalence rates of FASD have been observed (May et al., 2007; Viljoen et al., 2005).

The research presented here was conducted within the context of a large and multifaceted study initiative on various different aspects of fetal alcohol spectrum disorders (FASD), including diagnosis, epidemiology, dysmorphology, genetics and prevention (depicted in Figure 1). These activities are undertaken in partnership with the Foundation for Alcohol Related Research (FARR). The work presented here is effectively concentrated on the study design for a disease gene discovery approach, by focusing on effective candidate gene selection, as well as SNP selection for experimental data generation required for a statistical analysis such as an association study.

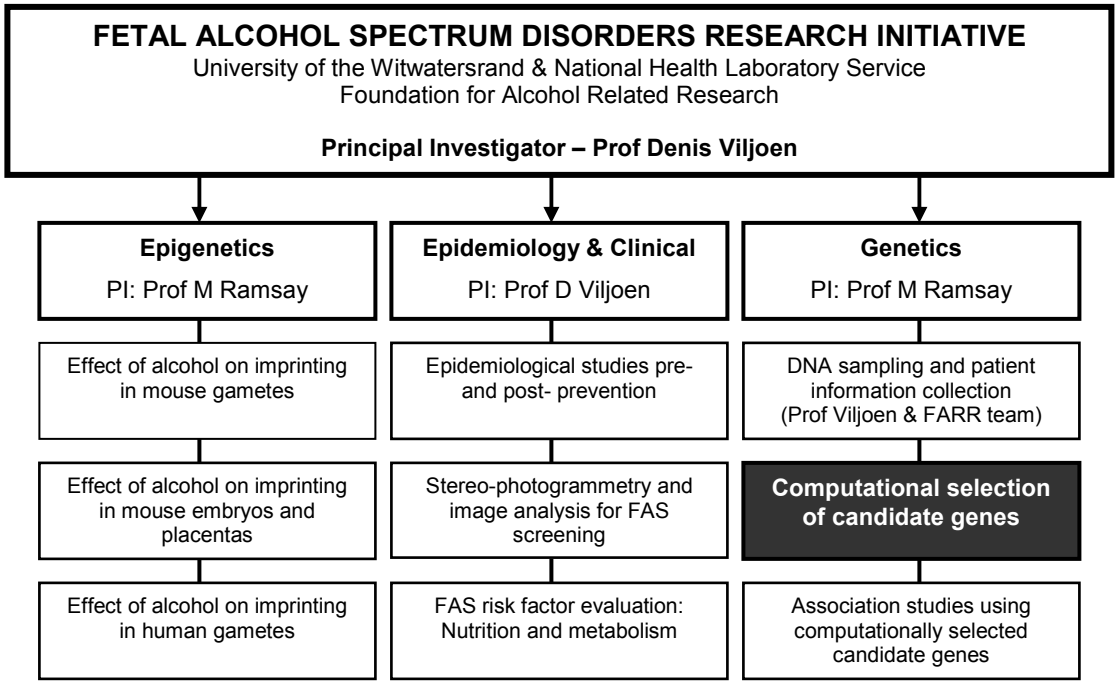

Figure 1: An outline of the different aspects of the FASD research initiative. The work presented in this thesis forms part of the genetics studies for the initiative (highlighted in the shaded block).

The towns of De Aar and Upington are the main focus in this study initiative (Figure 2). Although these two towns are both located in the Northern Cape province of South Africa, they are geographically distant (approximately 500 km apart). Upington is based on the banks of the Orange River, and is a prominent fruit and wine-growing town in South Africa. De Aar harboured the second largest railway junction in South Africa, until operations were significantly reduced during the latter half of the last century, resulting in an exceptionally high unemployment rates in the community. The fraught psychological and social circumstances largely contributed to the high levels of alcohol abuse in the community. The high FASD rates in Upington are thought to illustrate the effects of the drinking patterns developed amongst communities within the wine-growing regions; alcohol consumption is exceptionally high and as a result FASD is endemic.

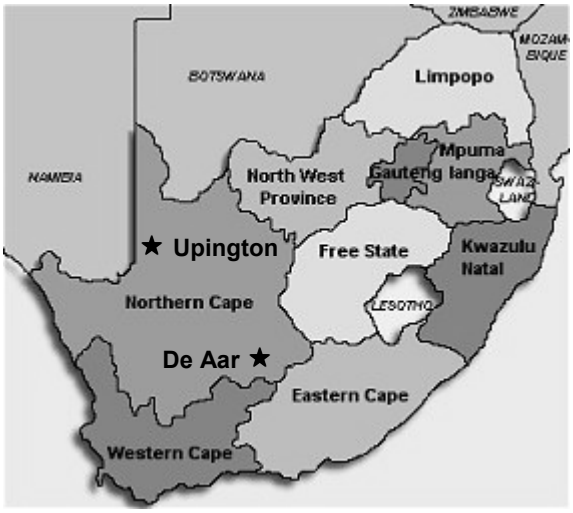

Figure 2: A map of South Africa indicating the geographical positioning of Upington and De Aar. Obtained from http://www.aboutsouthafrica.com/provinces.html

Over the years, strict guidelines have been set for complex disease association studies, which include a large sample size, highly significant *P* values, and verification of findings through replication in an independent sample. In addition associations should preferably be observed in both a family-based and population-based study (Todd, 2006). One of the goals of the FASD research initiative is to collect an adequately sized sample of affected participants and their family members, as well as a matched control group from the same community, to obtain the power required to make reliable conclusions from an association analysis. The goal is to use the information generated from this body of work to identify biologically plausible candidate genes for FASD, for such a prospective association study. *Chapter One* is a comprehensive overview of the aspects of study designs when

investigating the genetic influences for a complex disease, as well as a summary of computational methods available for candidate gene selection. The chapter also includes an outline of important features of FASD, particularly focusing on the influence of genetics on this spectrum of syndromes.

The prioritization method described here is based on a simulation of a researcher's approach to selecting candidate disease genes. In this process, information sources in the public domain are mined for candidate genes that exhibit characteristics relevant to disease phenotype. Consequently candidate genes were prioritized based on a binary filtering process. Secondly, the approach was used to select candidate genes for a disease with known genetic aetiology, to assess the precision with which putative disease genes are selected. For this study, X-linked mental retardation was selected as the test disease. *Chapter Two and Three* contain the description of this process for FASD and XLMR respectively.

The final section of this thesis is focused on selecting SNPs for association analysis, based on evaluating the putative effect of the genetic variants on protein function and gene expression (*Chapter Four*).

Note that throughout the text the American spelling of *fetal* will be used (as in fetal alcohol syndrome), as this is the generally accepted medical spelling. In general, *fetal* is the accepted spelling especially in the scientific community although *foetal* is still used in Commonwealth countries (http://en.wiktionary.org/wiki/fetal). It is internationally accepted that the medical spelling is the most appropriate phrase to use to access studies and international resources, as stipulated on the National organization on FAS-UK (NOFAS-UK) website *(*http://www.nofas-uk.org/). Other than this exception, South African spelling guidelines are followed throughout the text.

Publications and Presentations

The following publications and presentations arose from this research:

PUBLICATIONS:

- Lombard Z., Tiffin N., Bajic V.B., Hide W., Ramsay M. (2007) Selection and prioritization of candidate genes for fetal alcohol syndrome – a computational approach. *BMC Genomics* **8**: 389

ORAL PRESENTATIONS:

- 12th South African Society of Human Genetics Congress, Golden Gate, Free State, South Africa (2007)
  - Lombard Z., Kruger A, Tiffin N, Hide W, Ramsay M. In-silico candidate gene and SNP selection for a fetal alcohol syndrome association study
  - Received Discovery Institute Clinical Excellence Award for Best Oral Presentation
- 1st Southern African Bioinformatics Workshop, University of the Witwatersrand, Johannesburg, South Africa (2007)
  - Lombard Z., Ramsay M, Hide W. A computational approach to candidate gene prioritisation using data-mining and clinically-informed binary filtering.
  - Published in meeting proceedings as *work in progress*
- ACGT Bioinformatics and Functional Genomics workshop, University of the Witwatersrand (2006)
  - Lombard Z., Ramsay M. Identification of candidate genes for fetal alcohol syndrome
- The combined congress of the Southern African Society of Human Genetics and the African Society of Human Genetic, Muldersdrift, Gauteng (2005)
  - Lombard Z, Hide W, Ramsay M. A computational strategy to identify candidate genes for fetal alcohol syndrome – focus on the developing brain
- Faculty of Health Sciences Research Day, University of the Witwatersrand (2004)
  - Lombard Z, Ramsay M. A computational strategy to identify candidate genes for fetal alcohol syndrome – focus on the developing brain

POSTER PRESENTATIONS:

* Faculty of Health Sciences Research Day, University of the Witwatersrand (2006)
    - Lombard Z., Kruger A, Tiffin N, Hide W, Ramsay M. Identifying putative candidate genes through data-mining – Fetal Alcohol Syndrome as a model
    - Awarded best poster presentation in Molecular and Cellular Biology and Evolution Sciences track
* The Biology of Genomes Meeting, Cold Spring Harbor Laboratory, New York (10-14 May 2006)
    - Lombard Z., Kruger A, Tiffin N, Hide W, Ramsay M. Identifying putative candidate genes through data-mining – Fetal Alcohol Syndrome as a model

Table of contents

List of Tables

CHAPTER 1 – INTRODUCTION

CHAPTER 2 - COMPUTATIONAL SELECTION AND PRIORITIZATION OF CANDIDATE GENES FOR FETAL ALCOHOL SPECTRUM DISORDERS

CHAPTER 3 – ASSESSMENT OF THE COMPUTATIONAL PRIORITIZATION METHOD – X-LINKED MENTAL RETARDTION

CHAPTER 4 – ANALYSIS OF GENETIC VARIATION IN *FGFR1* – A CANDIDATE GENE FOR FETAL ALCOHOL SPECTRUM DISORDERS

List of Figures

CHAPTER 1 – INTRODUCTION

CHAPTER 2 - COMPUTATIONAL SELECTION AND PRIORITIZATION OF CANDIDATE GENES FOR FETAL ALCOHOL SPECTRUM DISORDERS

CHAPTER 3 – ASSESSMENT OF THE COMPUTATIONAL PRIORITIZATION METHOD – X-LINKED MENTAL RETARDTION

CHAPTER 4 – ANALYSIS OF GENETIC VARIATION IN *FGFR1* – A CANDIDATE GENE FOR FETAL ALCOHOL SPECTRUM DISORDERS

Abbreviations

˚C	degrees Celsius
ADH	alcohol dehydrogenase
ARBD	alcohol-related birth defects
ARND	alcohol-related neurodevelopmental disorder
BAC	blood alcohol concentration
BMI	body mass index
bp	base pairs
CAM	cell adhesion molecules
CCND1	G1/S-specific cyclin-D1
CHiP	Chromatin immunoprecipitation
CNS	central nervous system
CNV	copy number variation
CFG	convergent functional genomics
DAVID	database for annotation, visualization and integrated discovery
DMR	differentially methylated region
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DDE	dragon disease explorer
DTFAM	dragon transcription factor association miner
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FARR	Foundation for Alcohol Related Research
FAS	fetal alcohol syndrome
FASD	fetal alcohol spectrum disorders
FGF	fibroblast growth factor
FGFR1	fibroblast growth factor receptor 1
FGFR2	Fibroblast growth factor receptor 2
FOXG1B	forkhead box G1B
GABA	gamma-aminobutyric acid
GFINDer	genome functional integrated discoverer
GNAS	GNAS complex locus
GO	gene ontology
HGNC	HUGO gene nomenclature committee
HMGB1	high mobility group protein B1
HOXA1	homeobox A1

HUGO	Human Genome Organization
HWE	Hardy-Weinberg equilibrium
ICR	imprinting control region
Ig	immunoglobulin
IGF1R	insulin-like growth factor I receptor
IOM	Institute of Medicine (of the National Academy of Sciences)
IQ	intelligence quotient
Kb	kilobase
LD	linkage disequilibrium
LOD	logarithm of the odds
MAPK	mitogen-activated protein kinase
MeSH	medical subject headings
MGD	mouse genome database
MSX1	Msh homeobox homolog 1
NMDA	*N*-methyl-D-aspartate
NS-XLMR	non-syndromic X-linked mental retardation
OMIM	Online Mendelian Inheritance in Man
ORI	over-representation index
PE	promoter element
PCR	polymerase chain reaction
rSNP	regulatory SNP
Shh	sonic hedgehog
SIFT	sorting intolerant from tolerant
SNP	single nucleotide polymorphism
STRING	search tool for the retrieval of interacting genes/proteins
SVM	support vector machine
S-XLMR	syndromic X-linked mental retardation
T_A	annealing temperature
TF	transcription factor
TGF-β	transforming growth factor β
TFBS	transcription factor binding sites
TS	Theiller Stage
UCSC	University of California, Santa Cruz
UTR	untranslated region
V/cm	Volts per centimetre
w/v	weight per volume
XLMR	X-linked mental retardation

Chapter 1

Introduction

1.1 DISEASE GENE IDENTIFICATION

The identification and characterization of genes and genetic variants that result in disease, or contribute to disease susceptibility, is a critical objective in medical research. To date, such findings have proven to contribute greatly to improvements in diagnosis, prognosis and therapy (O'Connor and Crystal, 2006), accentuating the need to pursue the identification of as yet unknown disease-related genetic factors.

Initially, disease gene discovery focused on rare diseases, that all seem to have a Mendelian pattern of inheritance. Once appropriate statistical methods such as linkage analysis and abundant polymorphic markers such as restriction fragment length polymorphisms were available, many successful gene discovery studies were performed – including the elucidation of the genetic cause for cystic fibrosis (Kerem et al., 1989; Riordan et al., 1989), Huntington disease (Gusella et al., 1983; Huntington's Disease Collaborative Research Group, 1993) and Duchenne muscular dystrophy (Koenig et al., 1987) among others.

The typical approach taken for disease gene discovery (Figure 1.1), once the phenotype was defined, entailed the identification of affected families, genotyping and linkage analysis. Subsequent fine mapping of the linked region was applied to narrow down the candidate region and reduce the number of putative candidate genes, whereafter mutation detection within candidate genes was done to uncover the genetic cause of the disorder (Botstein and Risch 2003; Haines and Pericak-Vance, 2006). This process is essentially linear and unambiguous and has successfully been employed in the identification of a vast number of genes causing human disease.

With the successful disease gene discovery for many single-gene disorders, the focus shifted to diseases with a complex, multifactorial aetiology. The search for the genetic causes of complex (and often common) diseases has proven more complicated than initially anticipated. This is largely due to the fact that initially the same disease gene discovery approaches were employed for complex diseases as for monogenic disorders, even though the nature of the genetic contribution is expected to be quite different. Whereas one or a few rare mutations result in a monogenic disorder, complex diseases are caused by multiple genetic variants in multiple genes, each on their own not producing disease, but rather having a cumulative effect, in conjunction with environmental factors, on the risk of developing such a disease.

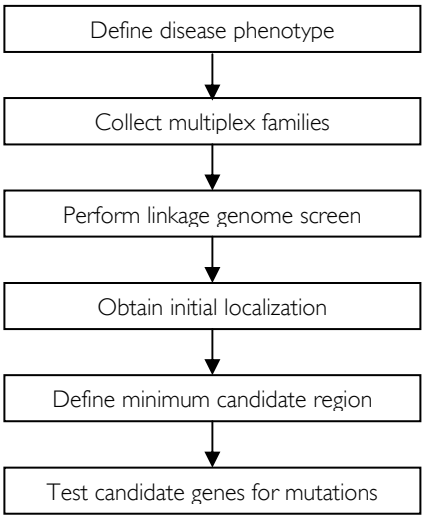

Figure 1.1: The typical approach taken to identify the genetic causes of single-gene disorders. Obtained from Haines and Pericak-Vance, (2006).

The steps taken to uncover the genetic risk factors for complex disease are therefore more intricate than for single-gene disorders, with different approaches being appropriate for different scenarios (Figure 1.2). It is important to note that there is no linearity to complex disease gene identification – rather the matters noted in Figure 1.2 are different points for consideration in the design of such an experiment, each having an impact on the other. As discussed in *Section 1.1.3*, linkage analysis is often not an appropriate choice for identifying the genetic causes in a complex disease. The major considerations when designing a study to identify genetic factors for a complex disease are outlined below.

1.1.1 Characterization of the phenotype

It is essential to ensure that the disease being studied is clinically well-characterized and that the clinical or phenotypic inclusion criteria are carefully assessed to exclude diseases that are likely to have a different molecular aetiology. Phenotype assessment must be done rigorously to prohibit false-positive or false-negative results being obtained during analysis. Many complex diseases have heterogeneous phenotypes, subtypes and age-related effects, emphasizing the importance of this step particularly when studying the genetic factors that contribute to complex disease (Tabor et al., 2002).

With monogenic disorders confirmation of a genetic contributor and identification of the pattern of inheritance is usually apparent when inspecting pedigree data, which is not the case with complex diseases. This implies that further analyses (such as segregation

analysis, twin- and adoption studies) are necessary to confirm whether genetics plays a role, or if the familial trend is due to other factors such as shared environment or biased ascertainment (Haines and Pericak-Vance, 2006).

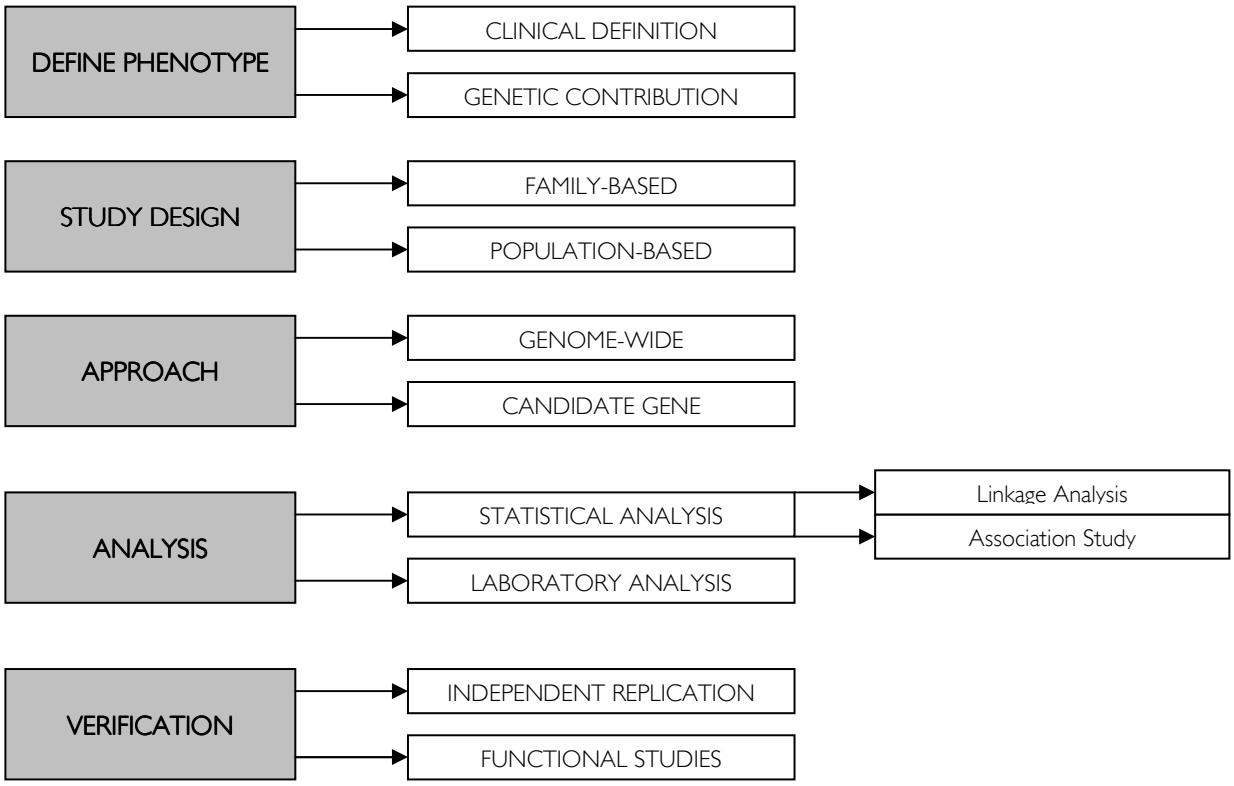

Figure 1.2: Key considerations when designing a study to identify genetic factors for a complex disease.

1.1.2 Study Design

The selection of a study design ties in closely with many of the other criteria that have to be regarded when designing a complex disease gene discovery experiment. Clearly characterizing the phenotype would give some indication of the most appropriate study design – for instance, it may not be possible to collect samples from family members of an individual with a late-onset disease, and a population-based design would therefore be more appropriate. Whether a family- or population based study design is chosen also depends on the statistical analysis that will be performed – there are different requirements and considerations for both linkage and association studies.

- **Family-based design**

Family-based designs are appropriate for both linkage- and association studies. For linkage analysis, extended families with multiple affected individuals or the affected sib-pair approach are most often used. Family-based association studies are based on the principle of testing the preferential transmission of a polymorphic allele from heterozygous parents to affected offspring through either the haplotype relative risk method (Falk and Rubinstein, 1987) or the transmission disequilibrium test (Spielman et al., 1994).

The disadvantages of family-based studies are that they are more time-consuming and costly than population-based studies, and generally have less power than case-control studies. However, these studies have the significant advantage that they are not affected by population stratification.

- **Population-based design**

The case-control approach involves genotyping individuals affected by a particular disease (i.e. cases) and a group of healthy, unrelated individuals (controls) and comparing the allele and genotype frequencies between these two groups (Mathew, 2001). It is important that case and control individuals are ascertained from the same study population and are matched for potential confounding factors (such as age and sex, if appropriate) to warrant credible findings. Collecting case-control samples is often the most viable approach in complex disease studies, and it has been shown that these studies have more statistical power to detect genetic risk than family-based studies (Daly and Day, 2001).

The major disadvantage of using this approach is that it is particularly prone to false-positive results (i.e. indicating that a genetic variant has a statistically significant influence on the disease when in fact it does not). This type of error most often occurs when using the case-control design due to population stratification. Population stratification refers to the presence of genetic subgroups within a study population, such as the case group (Figure 1.3). Alleles that have a different frequency in the genetic subgroup within the cases compared to that observed in the general population would then result in false-positive associations with that allele. (Balding, 2006; Freedman et al., 2004; Koller et al., 2004).

Population stratification can be compensated for by careful matching of cases and controls and the implementation of methods that control for stratification (discussed in *Section 1.1.3*).

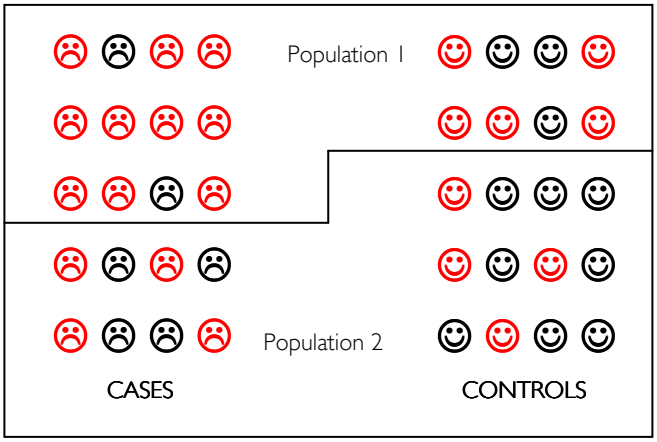

Figure 1.3: Population stratification. Population stratification arises due to the presence of genetic subgroups within the study population. In this scenario it would seem that the red allelic variant is associated with disease (it is more prevalent among the cases than the controls). This is however only due to the fact that the red allele is more prevalent in Population 1 than Population 2, and there are more representatives of that subgroup within the case group. Adapted from Balding (2006).

It is evident that both the family-based approach and the population-based approach have clear advantages and disadvantages. Therefore, it is accepted that these two approaches should rather be viewed as complementary designs, and are often both implemented in studies investigating the genetics of complex diseases.

1.1.3 Analysis

Although statistical analysis would only be done once some genotyping data has been generated, it is an essential matter that has to be considered when deciding on the study design. The statistical approach chosen to evaluate the disease-gene correlation impacts on the study design and approach selected.

- **Linkage analysis**

Genetic linkage refers to the co-inheritance of loci or alleles. Genetic loci that are physically close have a smaller chance to be separated during meiosis and crossing-over events, and are therefore considered to be linked. This observation led to the hypothesis that crossover frequency might indicate the distance separating genes on the chromosome, i.e. by calculating the number of recombinants it is possible to estimate the distance between genes. It is assumed that the greater the frequency of recombination between two genetic markers, the farther apart they can be assumed to be. The recombination frequency (θ) is a

measurement of recombination events and is a gauge of genetic linkage. $\theta = 0.50$ indicates that two loci are segregating independently. Significantly lower values are indicative of linkage.

Linkage analysis involves the genome-wide screening of DNA markers in families with multiple affected relatives to estimate the degree of linkage between a disease and a marker based on the genotypes observed in affected relatives (Melvin and Speer, 2006; Weir et al., 2006). In the presence of linkage it is expected that affected family members will inherit the disease-associated allele more often than would be predicted by the Mendelian principle of independent assortment. Markers that are physically close to this disease gene will be transmitted along with the disease allele, and therefore markers showing the strongest correlation and least recombination with the disease phenotype are assumed to be closest to the disease locus (Botstein and Risch, 2003).

It is necessary to statistically evaluate whether the correlation observed is unlikely to be a chance observation, confirming linkage. The most efficient statistic for evaluating this is the logarithm of the odds (LOD) score. A calculation is performed to assess the overall likelihood that the loci are linked ($\theta = 0$) or not linked ($\theta = 0.5$). The ratio of these two likelihoods gives the odds of linkage. Calculating the logarithm of this ratio gives the LOD score. A LOD score higher than three is usually accepted as a true indication of statistically significant linkage (Hoh and Ott, 2004).

Standard LOD score assessments are called parametric linkage analysis because it requires a precise genetic model, detailing the mode of inheritance, gene frequencies and penetrance of each genotype. This information is often not available for complex diseases, necessitating a non-parametric (or model free) approach (Dawn et al., 2005; Strauch et al., 2003).

Linkage analysis, although being both systematic and comprehensive, has one major drawback in that it has a relatively low power to detect markers exhibiting a moderate to small influence on disease susceptibility. With complex diseases, it is typically the case that various genes and genetic variants contribute to disease risk, with each gene having a small, often cumulative effect. Consequently linkage analysis is often not the appropriate choice to detect such genetic contributors. Furthermore locus heterogeneity, poor phenotype dissection and genotyping errors can have a negative influence on the statistical validity of this test. This consideration together with the fact that association analyses are

generally model-free have prompted many researchers to favour the association study to identify the genetic factors underlying complex disease risk (Botstein and Risch, 2003; Goldgar, 2001).

- **Association analysis**

Association studies compare allele frequencies between affected individuals and controls to assess the contribution of genetic variants to phenotypes. Association refers to the co-occurrence of an allele, genotype or haplotype with a disease trait, more frequently than can be readily explained by chance. The observation of a genotype-phenotype association could be the result of the variant having a direct functional impact, or by allelic association between the variant and another risk allele due to linkage disequilibrium (LD) (Mathew, 2001). There are different statistical approaches to association testing, including single variant and multipoint tests.

The simplest association analysis would be an allele-count test to evaluate the association of a single variant, such as a single nucleotide polymorphism (SNP) in a case-control study. A 2x3 table is constructed containing the SNP genotype counts for cases and controls, and either a Pearson χ^2 test or a Fisher exact test is performed to assess association. The Fisher exact test is the preferred test statistic if low genotype counts are a concern, as the χ^2 test's approximation is poor in such a scenario (Balding, 2006). Other options for evaluating single variant association are the Cochran-Armitage test (Armitage, 1955), linear regression and logistic regression (Balding, 2006).

For most common diseases with heritable components it is expected that multiple genetic variants in various genes influence susceptibility, and that much of the effect may be caused by interactions (epistasis) among multiple variants or environmental conditions (Tabor et al., 2002). Typically, many SNPs will be genotyped in a candidate gene, or high-density SNPs will be genotyped in a whole genome association study. Single-locus tests may lack power to detect association with disease, as each individual SNP provides relatively little information about LD with the disease variant; and corrections for multiple testing is likely to be conservative. Greater power is expected by joint analysis of all markers in the same gene or region simultaneously by considering multi-locus models of association. Logistic regression (Cordell and Clayton, 2002) and haplotype-based methods (Tzeng et al., 2003) are two accepted multi-locus association analyses. Recently, Bayesian

multilocus association methods have also been proposed for such analysis (Albrechtsen et al., 2007).

- **Additional statistical analyses**

Another consideration is the implementation of additional statistical analyses to ensure data quality and reliability.

Hardy-Weinberg Equilibrium (HWE)

Deviation from HWE should be assessed, as this could be indicative of genotyping error, inbreeding and population stratification. HWE can be assessed with a goodness-of-fit χ^2-test for biallelic markers, whereas accurate determination of HWE for markers with multiple alleles is achieved with the Fisher's exact test (Louis and Dempster, 1987) or a Markov Chain Monte Carlo method (Guo and Thompson, 1992).

Deviation from HWE can also be due to the presence of a common deletion polymorphism (Conrad et al., 2006) or a segmental duplication (Bailey and Eichler, 2006), and these scenarios should be considered before drawing conclusions from the results obtained for testing HWE.

Genomic control

Genomic control is an approach used to control for false-positive findings due to population stratification. Genomic control is based on the premise that null markers are genotyped genome-wide to estimate the confounding effect of population stratification and then removing the measured effect from the association test statistic. Genomic control can be applied to case–control studies using biallelic, multiallelic markers, or haplotypes (Devlin and Roeder, 1999; Devlin et al., 2001).

Corrections for multiple testing

When searching for the genetic variants that influence susceptibility to complex disorders, it is customary to test many variants in many genes, or even to genotype a genome-wide selection of genetic variants. This gives rise to the issue of multiple hypotheses testing where many statistical tests are performed on the same sample, resulting in inflated false-positive findings (Laird and Lange, 2006; Nyholt, 2001). The implication is that the

generally accepted significance level of α = 0.05 becomes an insufficient measure of significant findings. It is expected that more stringent criteria should be met to decrease the risk of declaring statistically significant effects when the findings are not significant but rather a result of inflated false-positives, due to multiple testing.

There is not consensus on which measure to implement (if any) to control for multiple testing error. Common techniques include the conservative Bonferroni correction and derivatives thereof (Gordi and Khamis, 2004) and procedures that monitor the false discovery rate (Benjamini and Hochberg, 1995).

1.1.4 Study approach

Linkage analysis is a hypothesis-independent strategy where the whole genome is systematically scanned, but is often not an appropriate approach complex disease gene mapping (discussed in *Section 1.1.3*).

For association studies, there are several study approaches that can be considered. The first group of study types are considered candidate approaches, as they focus on either a candidate polymorphism, candidate genes or the fine mapping of a candidate regions identified through linkage analysis. Alternatively, a whole-genome approach can be taken where the goal is to identify common causal variants throughout the genome using evenly spaced markers.

- **Candidate gene approach**

Traditionally association studies were restricted to the investigation of regions identified through linkage analysis and the testing of biologically relevant candidate genes within these regions – known as candidate gene studies. However, the main challenge when using this approach is the selection of suitable candidates to test, especially for diseases with poorly understood aetiology (Lohmueller et al., 2003; Tabor et al., 2002).

- **Whole-genome approach**

Whole-genome association studies – which entail the selection of a genome-wide set of genetic variants (usually SNPs) being typed to assess association – have in recent years become a viable approach for two main reasons. Firstly, the development of new high-throughput genotyping technologies (Tindall et al., 2007) (Sun and Guo, 2006) (Chen et al., 2003; Ji et al., 2004; Kim and Misra, 2007; Tost and Gut, 2005) ensured that the typing of

such a large selection of genetic variants is feasible. Secondely, the availability of data describing the common patterns of DNA sequence variation in the human genome such as the HapMap project (Frazer et al., 2007; International HapMap Consortium, 2003; International HapMap Consortium, 2005) enabled the implementation of whole-genome association studies.

The obvious advantage of this approach is that no prior knowledge regarding the underlying biology of the disease is needed. Recently such studies have yielded notable success for diseases such as Crohn disease (Mathew, 2008), type II diabetes (Salonen et al., 2007), obesity-related factors (Scuteri et al., 2007) and age-related macular degeneration (Klein et al., 2005). Most notable is a large-scale whole-genome association study by the Wellcome trust case control consortium that examined seven common diseases in 14000 cases and 3000 controls (The Wellcome trust case control consortium, 2007). This joint study identified 24 independent statistically significant association signals, highlighting the power of whole-genome studies when appropriate measures are implemented.

Since extended families are typically not used in whole-genome association studies, the region surrounding a marker that is shared identical-by-descent will be much smaller than the shared region for related individuals, due to the many more generations from the most recent common ancestor (Balding, 2006). The implication hereof is that a markedly higher marker density is needed than for linkage analysis, to adequately cover the whole genome. Typically a whole-genome association study would require genotyping in excess of 300 000 well-chosen SNPs in a Caucasian population, and even more for an African population due to the greater genetic diversity observed in these populations. This has a considerable impact on the cost of whole-genome association studies. Furthermore, the platforms needed to perform whole-genome association studies are at present not yet readily available in some countries and remain particularly expensive, inhibiting the use of this approach.

Estimates of human genetic variation suggest that even for those diseases where common genetic variants have been found, most genetic variation is still to be uncovered (Iles, 2008). Therefore, a further possible pitfall of the whole-genome association approach is that these studies are biased towards finding common variants, and that a class of variants that is too rare to be captured by these studies, but not sufficiently high risk to be captured by linkage analysis, may elude finding.Future novel approaches will be necessary to

elucidate these genetic factors of common disease, most probably utilising bioinformatics-based methods to identify candidate genes (Iles, 2008).

1.1.5 Verification

One of the concerning features of complex disease gene discovery is the significant number of studies reporting disease-gene associations that can not be replicated (Ioannidis et al., 2001; Lohmueller et al., 2003), casting doubt over the reliability of these findings. Recently reputable scientific journals have set gold standards for publishing such results (Nature Genetics Editorial, 1999; Nature Genetics Editorial, 2005), setting the guidelines for appropriate study design. These include large sample size, highly significant *P* values, and verification of findings through replication in an independent sample. In addition, the study should provide evidence that serves as justification for the association, such as in vitro cell-based studies or evidence from animal model systems. Finally, associations should preferably be observed in both a family-based and population-based study (Todd, 2006).

1.2 COMPUTATIONAL DISEASE GENE IDENTIFICATION

Recently, many computational candidate gene selection and -prioritization methods have been developed (Adie et al., 2006; Aerts et al., 2006; Franke et al., 2006; Freudenberg and Propping 2002; George et al., 2006; Kent et al., 2005; Lopez-Bigas and Ouzounis 2004; Perez-Iratxeta et al., 2005; Tiffin et al., 2005; Turner et al., 2003; van Driel et al., 2005). These tools aim to identify and prioritize putative disease genes by modelling specific characteristics of known disease genes, or by focusing on known disease features (such as gene expression profiles or phenotype).

1.2.1 Rationale for computational disease gene identification

Even though the traditional methods of disease gene identification have been effectively employed in identifying the genes underlying monogenic Mendelian disorders, the same success has not been achieved for complex diseases. Linkage analysis in particular has resulted in limited success due to the weak genotype-phenotype association which is a hallmark of multi-factorial disorders (Botstein and Risch, 2003). Association studies on the other hand have the main obstacle of selecting biologically relevant candidate genes to investigate, especially for diseases with poorly understood molecular aetiology. However,

apart from this obstacle, the candidate gene approach remains the most practical and frequently employed approach in disease gene investigation for complex disorders, justifying the need for approaches that aid candidate gene selection and prioritization.

Computational disease gene prediction methods are typically based on one of two aspects – gene properties and disease properties. Most tools available today employ either one of these or both aspects in their methodology, and are discussed below.

1.2.2 Approaches based on gene properties

- **Methodology**

It has been suggested that the genes underlying human hereditary disease share certain distinctive, sequence-based features, such as highly specific expression patterns (Bortoluzzi et al., 2003), distinctive sequence characteristics (Adie et al., 2005; Lopez-Bigas et al., 2006) and significantly differing mutation rates over evolutionary time as compared to non-disease genes (Smith and Eyre-Walker, 2003).

Some discrepant results regarding disease-gene and non-disease gene properties have been published. Whereas Smith and Eyre-Walker (2003) showed that disease genes have higher mutation rates over evolutionary time than other genes, Lopez-Bigas and Ouzounis (2004) and Adie et al. (2005) refuted these findings by showing that disease genes tend to be more highly conserved with a broader phylogenetic extent. The apparent discrepancy is clarified by Tu et al. (2006) who resolved the issue by classifying housekeeping genes as a category apart from other non-disease genes. By making this distinction the above-mentioned conflicting results were resolved by showing that the seemingly high mutation rates in disease genes (Smith and Eyre-Walker, 2003) were primarily due to the low mutation rates of the housekeeping genes within the non-disease group. Once the housekeeping genes were categorized separately, this observation diminished. Similarly the higher conservation level of disease genes (Lopez-Bigas and Ouzounis, 2004) was due to the faster evolution of non-housekeeping genes as opposed to house-keeping genes.

These differences can be harnessed to predict human disease genes, but have been shown to be quite non-specific, and may result in findings that are too extensive to be practically useful in pinpointing causative genes for complex diseases.

- **Examples**

The tools PROSPECTR (Adie et al., 2005) and DGP (Lopez-Bigas and Ouzounis, 2004) are examples of the computational methods available that predict candidate genes based on sequence properties.

The decision-tree computer learning algorithm of DGP is based on the hypothesis that genes involved in hereditary disease have some distinct sequence properties in common which render them more susceptible to mutations causing genetic disorders. DGP specifically evaluates protein length, degree of conservation, phylogeny and paralogy patterns to predict novel disease genes. When evaluated on a test set, 70% of the disease genes in the test set were predicted correctly with 67% precision (Lopez-Bigas and Ouzounis, 2004). However, when tested on the whole genome, this method classifies approximately 44% of the genome as probable disease genes. DGP is available for online use at *http://maine.ebi.ac.uk:8000/services/dgp*.

PROSPECTR is an alternating decision tree algorithm also based on disease gene sequence properties such as gene length, protein length and the percent identity of homologues. PROSPECTR was tested on a training set of genes consisting of 1084 genes known to be linked to disease and on a control set consisting of 1084 randomly selected genes not known to be involved in disease. It was found that 70% of the disease genes were correctly identified as such, whereas 43% of control genes were classified as false positives. PROSPECTR had a high false-positive rate when tested on the whole genome, classifying approximately 44% of all genes as likely to be involved in disease (Adie et al., 2005). PROSPECTR is accessible at *http://www.genetics.med.ed.ac.uk/prospectr*.

1.2.3 Approaches based on disease properties

- **Methodology**

In light of the non-specific results obtained using approaches based on sequence characteristics, prediction tools focusing on disease properties have been viewed more favourably. It is thought that similar phenotypes may be influenced by similar genotypes (Oti and Brunner, 2007), and therefore that comparable disease-associated phenotypic traits will have some disease genes in common. Similarly, it is assumed that genes that influence the same phenotype may be functionally correlated. Furthermore, it can be hypothesized that species sharing a phenotype are likely to have orthologous genes in the

concerned biological process, implying that genotype-phenotype correlations can be extrapolated across species. Therefore a variety of data sources have been employed in these computational methods and include the biomedical literature, functional annotations, gene expression data and data from animal model databases.

Computation methods based on this theory rely greatly on the annotation of genes, implying that these methods may falter in the face of poor or incomplete annotation. Furthermore, these methods can only be benchmarked on diseases with known genetic causes, which in most cases are monogenic disorders, implying that their effectiveness may be overestimated.

- **Examples**

A number of bioinformatics tools are available to select putative candidate genes based on disease properties. Examples of these tools are discussed in broad categories below based on the main data source that it utilizes. Additional tools are summarised in Table 1.1.

Scientific literature

The scientific literature remains the most comprehensive source of knowledge concerning disease and the genes involved, but the vastness thereof implies that effective extraction of knowledge is complicated (Korbel et al., 2005). The necessity therefore exists to devise methods that allow the rapid review of the available literature. Text-mining has become a popular approach employed in disease gene prediction, and is integral to the tools BITOLA (Hristovski et al., 2005) and a method that employs an integrative literature- and data mining approach to select candidate genes (Tiffin et al., 2005), amongst others.

BITOLA (Hristovski et al., 2005) is an interactive online data-mining tool available in two versions – CGI-BIN and as a Java applet (*http://www.mf.uni-lj.si/bitola/*). The goal of this system is to discover novel associations between diseases and genes, by mining the bibliographic database MEDLINE. This is achieved through the implementation of a unique discovery algorithm that connects a starting concept (disease name) with user-specified disease characteristics (such as pathological functions, symptoms, etc.) according to the literature. Genes that are related to the disease characteristics are then identified, and cross-matched to the chromosomal location linked to the disease. Identified candidate genes are also ranked by BITOLA using a heuristic ranking function (Hristovski et al., 2005). Being an automated system BITOLA does present some difficulty concerning

literature interpretation and terminology ambiguity, that may influence the accuracy of the results obtained.

The method described by Tiffin et al. (2005) is a text-mining technique, but also uses gene expression data by means of the eVOC ontology. The eVOC ontology is a controlled vocabulary used to describe the sample source of cDNA, SAGE libraries and labelled target cDNAs for microarray experiments. eVOC contains four orthogonal ontologies - anatomical system, cell type, pathology and developmental stage (Kelso et al., 2003). The method extracts eVOC terms from abstracts, which are subsequently ranked by calculating a ranking score for each associated eVOC term, according to the frequency of association and the frequency of annotation of the eVOC term. The four top-scoring eVOC terms are selected from the ranked list, and compared with eVOC terms annotated to genes within the Ensembl database to select putative candidates. Allowance is made in the system for one mismatch, implying that genes selected as candidates are those annotated with at least three of the four top-scoring eVOC terms. This approach was tested on a subset of genes (training database of 417 genes) representative of those that might be selected by a linkage analysis study, and not the full complement of genes in the Ensembl database. Although not available as an online tool, the scripts and test dataset can be downloaded from *http://www.sanbi.ac.za/tiffin_et_al/*.

The level of complexity in the scientific literature and the understandable constraints of natural language processing are two major stumbling blocks in effective text-mining, and contribute to most of these techniques being rather unspecific and resulting in high false-positive rates.

Gene expression data

Gene expression is often dysregulated in tissues affected by disease, and it is therefore a useful indicator to consider in the search for complex disease candidates. GeneSeeker and TOM are examples of computational tools that focus on gene expression profiles to identify and prioritize candidate genes.

GeneSeeker (van Driel et al., 2003; van Driel et al., 2005) is a modular web tool that uses known positional information, as well as user-defined expression and phenotype terms for candidate gene evaluation. The information provided is used to search the GDB human genome database and MIMMAP (a reformatted version of the Online Mendelian Inheritance in Man (OMIM) gene mapping information) for genes in a specified

chromosome location, while the mouse genome database (MGD) is queried for mouse genes in the homologous regions. Various gene expression and phenotypic databases are then searched for all genes that match the given expression terms. This results in a swift overview of the biological relevance of candidate genes in the region of interest. GeneSeeker is available via the web interface *http://www.cmbi.kun.nl/GeneSeeker/*.

TOM (Rossi et al., 2006) is an automated pipeline for candidate gene prediction, based on microarray gene expression data and functional annotation. The algorithm employed in TOM requires that at least one other gene responsible for the disease and the linked chromosomal region be known, or else, that at least two genetic regions have been linked to the disease. In the first case where at least one disease gene is known, the algorithm extracts the genes within the linkage region that has the highest likelihood of being functionally related to the known gene. In the absence of a known disease gene, the algorithm extracts the genes annotated in the given linked regions and searches for those that have similar expression- or functional profiles. The algorithm is available at *http://www-micrel.deis.unibo.it/~tom/*.

Tools that focus on gene expression patterns have some drawbacks when implemented in identifying candidate genes for complex disease. The availability of expression data, and more specifically expression data for certain tissue types, remains sparse in public databases (Kelso et al., 2003), complicating the efficient implementation of these tools. It should also be considered that for many complex diseases multiple systems and tissues may be involved, complicating the choice of target tissue to focus on, and limiting the appropriateness of this approach.

Functional annotation

It is assumed that genes that influence the same phenotype may be functionally correlated, as it has been shown for many diseases that similar phenotypes result from similar genotypes. Freudenberg and Propping (Freudenberg and Propping, 2002) systematically analyzed the OMIM database to identify relationships between gene function and phenotype, showing that a strong correlation exists between similar phenotypes and a number of gene functions. Various tools that focus on functional annotation data sources exist.

POCUS (Turner et al., 2003) assesses the over-representation of functional annotation terms for genes in loci associated with a disease, and requires no prior knowledge of the

disease of interest other than the identified associated genomic locus. POCUS functions on the premise that genes underlying complex diseases tend to share functional annotation, and assumes that a gene is a putative candidate based on the observation that two or more of the genes within the linked regions share some aspect of their expression pattern or of the function of the encoded protein. This may be problematic when applied to complex diseases, as the molecular mechanisms underlying these diseases may not always be functionally related. POCUS is available at *http://www.hgu.mrc.ac.uk/Users/Colin.Semple/*.

Convergent Functional Genomics (CFG) is an approach used to identify and prioritize candidate genes, it relies on the cross-matching of animal model gene expression data and functional annotation with human genetic linkage data (Bertsch et al., 2005; Rodd et al., 2007). This approach uses a Bayesian-like methodology of reducing uncertainty through the combination of multiple independent lines of evidence, each by itself lacking sufficient power to confirm that a gene is a putative candidate gene, to produce a short list of high probability candidate genes (Rodd et al., 2007).

1.2.4 Integrative approaches

- **Methodology**

All the techniques discussed thus far have both advantages and pitfalls, which leads one to consider that an integrative approach that harnesses the benefits of each technique would be beneficial in the elucidation of the genetic causes in complex diseases. Tiffin et al. (2006) recently surveyed some of the methods for computational disease gene identification and concluded that using the methods in concert was more successful in prioritizing candidate genes for disease, than when each was used alone. This review additionally showed that using existing computational methods in concert highlighted potential candidates that are selected by a subset of methods and are missed by the other methods, depending on the type of data examined. This observation gives further evidence that the inclusion of more data sources may positively aid disease gene discovery.

- **Examples**

Below a few techniques that use multiple data sources to assess gene candidacy are discussed, with more examples in Table 1.1.

The Genome Functional Integrated Discoverer (GFINDer) (Masseroli et al., 2004) is an automated web server tool intended to identify the functional annotation terms enriched within a user provided gene list, and therefore functionally classifies the list according to several functional categories, which include Gene Ontology (GO) classes of biological processes, cellular components and molecular functions, KEGG biochemical pathways, PFAM protein domains and OMIM diseases (Masseroli et al., 2005; Masseroli et al., 2004). GFINDer is therefore most appropriately applied to biologically interpreting a sub-list of genes, rather than selecting candidate genes from the whole genome. GFINDer is available at *http://www.medinfopoli.polimi.it/GFINDer/*.

SUSPECTS (*http://www.genetics.med.ed.ac.uk/suspects/*) is an extension of PROSPECTR (described in *Section 1.2.7*) that incorporates co-expression, shared protein domains and functional annotation data in addition to sequence-based evidence to more effectively select candidate genes (Adie et al., 2006). A submitted gene list is evaluated on these features, and compared to genes known to be linked to the disease of interest, in order to rank them according to the likelihood that they are involved in a particular disorder.

Endeavour is maybe the most flexible and integrative of all tools discussed thus far. This freely accessible, interactive and flexible software prioritizes candidate genes based on the premise that genes involved in the same disease share annotations and other characteristics (Aerts et al., 2006). A list of candidate genes are evaluated and ranked according to their similarity with known 'training' genes pertinent to the disease. Training genes are selected based on a GO term, a KEGG pathway identifier or an OMIM disease name. In order to evaluate the parallels between the gene list and the training set, substantially more data types are accessed, including literature, functional annotation, microarray expression, EST expression, protein domains, protein-protein interactions, pathway membership, *cis*-regulatory modules, transcriptional motifs and sequence similarity. Furthermore, Endeavour offers the flexibility of including additional, user-defined sources, optimizing the tool's effectiveness by incorporating expert knowledge. However, in the case of diseases with no linked genomic region and where the disease mechanisms are not entirely clear (which is often the case for many complex diseases) choosing a training set is not feasible, prohibiting the use of this method in such cases. Endeavour is available for download from *http://www.esat.kuleuven.be/endeavour*.

A major drawback of most of the methods described above is the assumption that chromosomal regions linked to the disease of interest have been identified, or that some prior

knowledge of genetic aetiology exists. In the absence of such knowledge, the usefulness of these tools diminishes. This particular study focuses on fetal alcohol spectrum disorders (FASD), which has a complex and poorly understood aetiology. Furthermore no linkage studies have been performed, signifying that no region of the genome has conclusively been linked to risk of FASD development. Therefore strong motivation exists for using novel computational tools to identify possible candidate genes for FASD development, as the study approach.

Table 1.1: A list of several available computational candidate gene selection tools. Modified from Oti and Brunner 2007.

Tool	Data source used	Online availability
DGP	Sequence	http://maine.ebi.ac.uk:8000/services/dgp
PROSPECTR	Sequence	http://www.genetics.med.ed.ac.uk/prospectr
BITOLA	Literature	http://www.mf.uni-lj.si/bitola/
Tiffin et al. (2005)	Expression & literature	http://www.sanbi.ac.za/tiffin_et_al/
Genes2Disease	Functional annotation &sequence	http://www.ogic.ca/projects/g2d_2/
GeneSeeker	Expression, literature & phenotype	http://www.cmbi.ru.nl/GeneSeeker/
TOM	Expression & functional annotation	http://www-micrel.deis.unibo.it/_tom/
POCUS	Functional annotation	http://www.hgu.mrc.ac.uk/Users/Colin.Semple/
CFG	Expression & functional annotation	~
GFINDer	Expression, functional annotation, phenotype , pathway data & protein domains	http://www.bioinformatics.polimi.it/GFINDer/
SUSPECTS	Functional annotation & sequence	http://www.genetics.med.ed.ac.uk/suspects/
Endeavour	Expression, functional annotation, literature, pathway data, phenotype, protein domains & other	http://www.esat.kuleuven.be/endeavour/
Prioritizer	Expression, functional annotation & protein-protein interaction	http://www.prioritizer.nl/
UCSC Genesorter	Expression, literature & sequence	http://www.genome.ucsc.edu/cgi-bin/hgNear

1.3 FETAL ALCOHOL SPECTRUM DISORDERS

Fetal alcohol spectrum disorders (FASD) is an umbrella term used to describe the range of disorders that result from in utero alcohol exposure (FASD terminology summit consensus statement, 7 April 2004). The term FASD is not intended for use as a clinical diagnosis, but rather encompasses the range of congenital abnormalities that can result from exposure to alcohol's teratogenic effect during prenatal development. These anomalies include prenatal and postnatal growth retardation, central nervous system (CNS) dysfunction, characteristic craniofacial malformation and other organ abnormalities (Clarren et al., 1978; Day et al., 1999; Sulik and Johnston 1983). Fetal alcohol syndrome (FAS) is the clinical description for children at the most severe end of the FASD spectrum, who display the full phenotype associated with in utero alcohol exposure. FASD encompasses a series of serious and permanent disorders, with severe physiological, psychological and social implications for the affected individual. Prenatal alcohol exposure is known as an entirely preventable cause of mental retardation (Abel, 1995), and FASD is a prominent public health problem in South Africa.

1.3.1 Clinical features

The teratogenic effects of alcohol have been recognized for centuries, but it wasn't fully described and recognized as a discrete disease entity until the seventies when Jones and his colleagues coined the term FAS (Jones and Smith, 1973; Jones et al., 1973) and published a comprehensive description of the malformations observed in children of mothers who consumed alcohol during pregnancy. The following three clinical characteristics constitute the main features of FASD:

Characteristic pattern of facial anomalies

Several craniofacial abnormalities are associated with FASD, including short palpebral fissures, a smooth or flattened filtrum, a thin vermilion border, epicanthic folds, a flattened nasal bridge and a short upturned nose, as well as micrognathia (Manning and Hoyme, 2007).

Evidence of growth retardation

Both pre- and postnatal growth deficiency is a hallmark of prenatal alcohol exposure. Children at or below the 10^{th} percentile on height, weight and head circumference charts are considered to be below the average height and weight for their age group (Hoyme et al., 2005; Manning and Hoyme, 2007).

CNS abnormalities

CNS dysfunction is the most severe and permanent consequence of in utero alcohol exposure and is a feature in most syndromes falling within the FASD spectrum. CNS damage can present as structural, neurological and functional deficit as summarised in Table 1.2.

Table1.2: The consequences of alcohol-induced CNS damage

Structural	Functional	Neurological
Microcephaly Change in size or shape of corpus callosum, cerebellum, or basal ganglia	Lack of impulse control Verbal processing problems Learning disabilities such as attention deficit Higher-level receptive & expressive language deficits Impairment in complex task performance (problem solving, planning & judgment)	Epilepsy Impaired fine motor skills Neuro-sensory hearing loss Poor gait and clumsiness Eye-hand coordination deficits

There are several other features that could also present in individuals exposed to alcohol in utero. Therefore a checklist of anomalies, such as that noted in Table 1.3 may be used to quantify features during the evaluation of patients (Hoyme et al., 2005; Manning and Hoyme, 2007). This dysmorphology scale is used to assess the level of malformation and not to assign a diagnosis.

1.3.2 Diagnosis

Since the description of FAS (Jones and Smith, 1973; Jones et al., 1973) several guidelines for diagnosing FASD have been published (Astley and Clarren, 2000; Center for Disease Control, 2004; Chudley et al., 2005; Stratton et al., 1996). The Institute of Medicine (IOM) of the National Academy of Sciences released the first standardized diagnosis system (Stratton et al., 1996), which at present, in conjunction with the Washington criteria (Astley and Clarren 2000), are the diagnostic systems most often used for FASD diagnoses.

Table 1.3: The scoring system used to assess dysmorphic features in children prenatally exposed to alcohol, as described by Hoyme et al. (2005) and Manning and Hoyme (2007). A specific score is assigned for each feature presented by the individual; whereafter a final score is summed.

Feature	Score
Height deficits (below 10th centile)	1
Weight deficits (below 10th centile)	2
Occipitofrontal circumference (below 10th centile)	3
Inner canthal distance(below 10th centile)	0
Palpebral fissure length (below 10th centile)	3
Strabismus	0
Ptosis	2
Epicanthal folds (non-racial)	1
Flat nasal bridge	1
Anteverted nares	2
Long philtrum	2
Smooth philtrum	3
Thin vermilion border of upper lip	3
Midfacial hypoplasia	2
‘‘Railroad track’’ ears	1
Prognathism	0
Cardiac murmur	0
Confirmed cardiac malformation	1
Decreased pronation/supination of elbow	2
Hypoplastic nails	0
Clinodactyly of fifth fingers	1
Camptodactyly	1
‘‘Hockey stick’’ palmar creases	1
Hirsutism	1
Attention-deficit/hyperactivity disorder	1
Fine motor dysfunction	1
Total possible dysmorphology score	**35**

Recently, clarifications of the 1996 IOM criteria for the diagnosis of FASD were published (Hoyme et al., 2005; Manning and Hoyme, 2007) to assist the practical application of these criteria in clinical practice.

These reports suggest the following six diagnoses to include the full spectrum of FASD:

- FAS with confirmed maternal alcohol exposure
- FAS without confirmed maternal alcohol exposure
- Partial FAS with confirmed maternal alcohol exposure
- Partial FAS without confirmed maternal alcohol exposure
- Alcohol-related birth defects (ARBD); and
- Alcohol-related neuro-developmental disorder (ARND)

Table 1.4 shows the six diagnostic categories and the criteria necessary to make a specific diagnosis.

Table 1.4: Proposed clarification of the IOM criteria for diagnosis of FASD. Obtained from Hoyme et al. (2005) and Manning and Hoyme (2007)

I. FAS WITH CONFIRMED MATERNAL ALCOHOL EXPOSURE
A. Confirmed maternal alcohol exposure B. Evidence of a characteristic pattern of minor facial anomalies, incl. two or more of the following 1. Short palpebral fissures (below 10th percentile) 2. Thin vermilion border of the upper lip 3. Smooth philtrum C. Evidence of prenatal and/or postnatal growth retardation 1. Height or weight below 10th percentile (corrected for racial norms, if possible) D. Evidence of deficient brain growth or abnormal morphogenesis, incl. one or more of the following: 1. Structural brain abnormalities 2. Head circumference below 10th percentile
II. FAS WITHOUT CONFIRMED MATERNAL ALCOHOL EXPOSURE
IB, IC, and ID, as above
III. PARTIAL FAS WITH CONFIRMED MATERNAL ALCOHOL
IA and IB, as above C. One of the following other characteristics 1. Evidence of prenatal and/or postnatal growth retardation (height or weight below 10th percentile) 2. Evidence of deficient brain growth or abnormal morphogenesis 3. Evidence of behavioural or cognitive abnormalities inconsistent with developmental level that cannot be explained by genetic predisposition, family background, or environment alone
IV. PARTIAL FAS WITHOUT CONFIRMED MATERNAL ALCOHOL EXPOSURE
IB and IIIC, as above
V. ARBD
IA and IB, as above C. Congenital structural defects in one or more of the following categories (2 or more for minor anomalies): 1. Cardiac: atrial septal defects, aberrant great vessels, ventricular septal defects, conotruncal heart defects 2. Skeletal: radioulnar synostosis, vertebral segmentation defects, large joint contractures, scoliosis 3. Renal: aplastic/hypoplastic/dysplastic kidneys, “horseshoe” kidneys/ureteral duplications 4. Eyes and ears: strabismus, ptosis, retinal vascular anomalies, optic nerve hypoplasia; conductive hearing loss, neurosensory hearing loss 5. Minor anomalies: hypoplastic nails, short fifth digits, clinodactyly of fifth fingers, pectus carinatum/excavatum, camptodactyly, “hockey stick” palmar creases, refractive errors, “railroad track” ears
VI. ARND
A. Confirmed maternal alcohol exposure B. At least 1 of the following 1. Evidence of deficient brain growth or abnormal morphogenesis 2. Evidence of behavioral or cognitive abnormalities inconsistent with developmental level that cannot be explained by genetic predisposition, family background, or environment alone

1.3.3 FASD in South Africa

Epidemiological studies in a South African community of mixed-ancestry (also known as the Coloured community) in the Western Cape Province have revealed an exceptionally high FASD rate, accentuating the gravity of FASD as a public health issue in South Africa. The range of prevalence rates reported in two different primary school cohorts from this community were 65.2-74.2 per 1 000 (Viljoen et al., 2005) and 68.0-89.2 per 1000 (May et al., 2007) respectively. This rate is alarmingly higher than the average observed for the developed world of 0.97 per 1000 live births (Abel, 1995). Preliminary studies in two Northern Cape Coloured communities propose similarly high rates, with suggested prevalence rates as high as 67-103 per 1000 school-entry aged children (DL Viljoen, personal communication). It is suspected that other communities in South Africa may have equally high rates; accentuating the need for an increase in the understanding of disease aetiology, and implementation of prevention programs.

1.3.4 Risk factors for FASD

Although alcohol consumption during pregnancy is the primary trigger for the presentation of FASD, the exact mechanisms for alcohol-induced teratogenic effects have not been fully clarified. The role of additional factors other than alcohol in FASD development is primarily supported by the observation that FASD does not occur in all children exposed to alcohol during the prenatal period (Bonthius et al., 2004). The array of alcohol-related birth defects that falls within the FASD spectrum, further suggests variation in the effect that alcohol has on individual fetuses.

The factors that influence FASD development can be divided into intrinsic factors – such as maternal and fetal genotype, and epigenetics; and extrinsic factors such as environmental causes and socio-economic status (Abel, 1995; Maier and West, 2001; Sokol et al., 1986; Warren and Foudin, 2001; West et al., 1990).

- **Extrinsic factors**

Environmental factors other than in utero alcohol exposure appear to play a big role in the effect of alcohol exposure. FASD has been associated with specific drinking patterns, body mass index (BMI) and lifelong and current nutrition, advanced maternal age, high gravidity and parity, unstable marital status, cigarette use, and use of other drugs (May et al., 2005; Viljoen et al., 2002). Episodic binge drinking that produces high blood alcohol concentration

(BAC) levels is a well known maternal risk factor that is known to cause severe damage to the CNS of the developing fetus. Furthermore, a dose- and time-dependant relationship has been observed, where exposure to higher concentrations of alcohol at critical developmental stages resulted in more severe anomalies (Sampson et al., 1997). Poor nutrition likely contributes to a higher risk rate for FASD in that malnourished individuals may less effectively metabolize alcohol, thus allowing more alcohol to cross the placenta and cause more fetal damage (Khaole and Li; 2000). Low socio-economic status, lower educational attainment and reported lower religiosity have also been linked to FASD risk. These factors contribute on a social level to create an environment conducive to alcohol abuse, even when pregnant.

- **Intrinsic factors**

It is interesting to note that a study by May et al. (May et al., 2005) observed a disparity in risk between South African women and American Indian women from the United States. Even though women in the United States sample report a high consumption of alcohol in a binge pattern of drinking, less detectable damage to the fetus was observed than among the South African women (May et al., 2004). This disparity indicates that genetic and epigenetic factors may also play a role in risk of developing FASD.

It has been shown that an organism's phenotype is not only determined by the parental DNA, but that epigenetic modifications could also be transferred from one generation to the next, ultimately influencing the phenotype. Epigenetics refer to changes in gene expression that take place without a change in the DNA sequence, due largely to molecular modifications to both DNA and chromatin, the most extensively investigated of which are DNA methylation patterns. DNA methylation refers to the endogenous process in a cell where a methyl group is added to the cytosine or adenosine bases in a stretch of DNA, resulting in gene-silencing (Jirtle and Skinner, 2007).

Genomic imprinting is an epigenetic mechanism, resulting in the preferential expression of either the paternal or maternal allele of certain genes (Surani, 1998). One of the characteristics of imprinted genes is differential allele-specific DNA methylation, which is usually localized to regions known as differentially methylated regions (DMRs) of which there are generally two varieties. Firstly, there are DMRs that acquire differential methylation during somatic development, in a tissue-specific manner (Reik and Walter, 2001). The second variety of DMRs are referred to as imprinting control regions (ICRs),

which are differentially methylated in all tissues, throughout development. ICRs are the principal regulators of gene expression within imprinted domains, often over large distances (Robertson et al., 2006).

The expression of many prokaryotic and eukaryotic genes is regulated through the methylation of DNA (Lim and van Oudenaarden, 2007). Animal studies have shown that in utero ethanol exposure inhibits global fetal DNA methylation (Garro et al., 1991; Valles et al., 1997). Since DNA methylation and imprinting play an important role in the regulation of gene expression during embryogenesis (Garro et al., 1991; Wagschal and Feil, 2006) and consequent development, ethanol-associated alterations in fetal DNA methylation may contribute to the developmental abnormalities seen in FASD.

Twin concordance studies (Hao et al., 2003), animal model systems (Hardy, 1999; Michelson et al., 2002) and the recent observation that paternal alcohol consumption may negatively influence fetal neuro-behavioural development and growth (Davies, 2003; Ehringer and Sikela, 2002) lend support to the notion of genetics playing a role in FASD development, and are discussed in more detail below.

1.3.5 Genetic risk factors for FASD

FASD can be considered to be a multi-factorial or complex disease, implying that the genetic factors underlying FASD, as well as the environmental risk factors, will be plentiful and the interactions between these factors intricate. There are two main lines of evidence that support the theory that genetics plays a role in FASD development – twin concordance studies and animal model systems.

- **Twin concordance studies**

An association between a variable genetic background and FASD development is primarily supported by the observation that FASD does not occur in all children exposed to alcohol during the prenatal period (Chaudhuri, 2000). This observation suggests that certain individuals may have a genetic predisposition to infliction of more severe damage by gestational alcohol exposure; and the varied phenotype observed in FASD may be a reflection of the varied susceptibility quotients in the genetic background of the individual. Streissguth and Dehaene (Streissguth and Dehaene, 1993) studied twin pairs with alcoholic mothers, and found the rate of concordance for FASD to be 100% for

monozygotic twins, whereas dizygotic twins showed only 64% concordance. Maternal alcohol consumption during pregnancy was confirmed; however only one of the twin pair developed FAS, whereas the other was unaffected.

- **Animal model studies**

Animal models have proven very useful in clarifying the prenatal effects of alcohol exposure, and elucidating the role of genetics in FASD development. Different animal model systems have been used, with the zebrafish being the earliest model to illustrate alcohol's teratogenic effects (Stockard, 1910). The zebrafish is still used to examine fundamental questions about the effects of embryonic exposure to ethanol on development (Bilotta et al., 2004).

Although many animal models for FAS exist, the mouse model seems to correlate best to the phenotypic effects of prenatal alcohol exposure observed in humans (Sulik, 2005). Sulik et al. (Sulik et al., 1981) illustrated that mice exposed to alcohol in utero showed microcephaly, microphthalmia with accompanying short palpebral fissures, and a long upper lip with a deficient philtrum, comparable to the phenotype observed in humans with FAS. The degree of severity of the alcohol-induced facial defects in the mice also varies widely, as is the case with FASD in humans (Sulik, 2005). Due to this extrapolation of experimental results from the animal model to the human phenotype, the FAS mouse model also lends itself to the exploration of genetics and prenatal alcohol risk.

Diallelic crosses performed on CBA, C3H, and C57BL mouse strains, showed that the extent of teratogenesis was found to be a function of the genetics of the mothers rather than the fetus (Chernoff, 1977). Further support for the role of the maternal genotype in FASD development is provided in a study by Gilliam et al. (Gilliam et al., 1997) where two inbred strains of mice with differing sensitivity for alcohol-induced birth defects were crossed. This study showed that fetuses with an alcohol-susceptible mother had an almost nine times higher rate of malformation than fetuses with an alcohol-resistant mother. However, a study by Ogawa et al. (Ogawa et al., 2005) showed that the fetal genotype may also contribute to FASD development, by illustrating that alcohol exposed embryos exhibited deficits that could only be attributed to genetic factors within the fetus. Several other studies have shown similar strain-related differences in the extent and pattern of

alcohol-induced malformations, as well as behavioral outcome (Boehm et al., 1997; Gilliam et al., 1997; Thomas et al., 2000; Thomas et al., 1998).

1.4 PUTATIVE GENETIC TARGETS OF ALCOHOL

This study will focus on identifying genetic targets of alcohol-induced damage during fetal development. A number of potential leads in understanding the genetics of FASD are derived from studies focusing on alcohol metabolism. These studies have generally focused on the alcohol dehydrogenase (ADH) enzyme family members and conflicting results have been obtained. Stoler et al. (2002) observed that the absence of the ADH1B*3 allele was protective for fetal outcome, in conflict with two other studies showing the presence of this allele to be protective (Jacobson et al., 2006; McCarver et al., 1997). The ADH1B*2 allele has been proposed to play a possible protective role, or to be a marker for protection in the South African mixed-ancestry population (Viljoen et al., 2001). However, the sample size for this association study was small, and results have not yet been replicated in other populations. Many other genes are likely to contribute towards the development of FASD and further investigation is required.

CNS dysfunction is the most severe and permanent consequence of maternal alcohol intake and is the only feature present in all other disorders in the spectrum of alcohol-related birth defects. It is therefore expected that pathways and systems within the CNS, and associated to its development will be prominent targets in the search for putative genetic factors for FASD. Alcohol has a complex teratogenic effect on the CNS, having many cellular targets and modes of toxicity (Goodlett and Horn, 2001). Two of the main functional systems affected by alcohol exposure, namely, neural apoptosis and cell-cell interaction, are discussed below.

1.4.1 Neural apoptosis

Investigating alcohol-induced neural apoptosis may elucidate part of the genetic component of FASD development, as apoptotic cell death entails the activation of a gene-directed program for cellular self-destruction (Clarren et al., 1978). Several studies have shown that alcohol suppresses neuronal activity, resulting in a pro-apoptotic environment in the developing brain (Galvan et al., 2003; Thiery, 2003). Alcohol-induced neural apoptosis has been observed throughout the developing CNS, including all levels of the spinal cord,

brain stem, cerebellum, midbrain and forebrain. Furthermore, alcohol has been observed to diminish neurons from various parts of the developing visual-, auditory- and memory systems of the developing brain (Thiery, 2003). This pro-apoptotic effect of alcohol provides a probable explanation for the long-term CNS dysfunction and diminished brain size associated with FASD.

Alcohol has an array of molecular pathway targets and modes of inducing apoptosis. Genes in these pathways may all be potential candidates for susceptibility to FASD development. They include:

- **Neurotransmitter receptors and intermediates**

Alcohol disrupts neurochemical-signalling pathways through interaction with both the neurotransmitter receptor and its substrates. The two major receptors directly affected by alcohol are ligand-gated ion- and voltage-dependant calcium channels. The gamma-aminobutyric acid (GABA) and *N*-methyl-D-aspartate (NMDA) molecules and their receptors are of particular interest, as alcohol exhibits both NMDA antagonist and GABA-mimetic properties, which subsequently lead to neural apoptosis (Thiery, 2003).

Blocking of NMDA glutamate receptors (antagonism) has been shown to substantially increase neural apoptosis. NMDA antagonism results in the developing CNS counteracting the inhibitory effect through increasing the number of NMDA receptors. With the elimination of alcohol from the fetal system, the sections of the CNS with elevated NMDA receptor levels, experience hyper-excitability, which leads to the activation of a cascade of intracellular events, and ultimately, to neuronal death (Hasty et al., 2001).

GABA acts as a neurotransmitter in ~20% of CNS synapses and is an essential inhibitory substance in the brain. Prenatal alcohol exposure has been shown to mimic or potentiate the action of GABA and its receptor. This GABA-mimetic effect results in sustained stimulation of GABAergic neurons, resulting in high levels of glutamate, which ultimately leads to neuronal death (Hsiao et al., 2004).

- **Oxidative stress and mitochondrial dysfunction**

Another mechanism by which alcohol induces apoptosis in the brain is by stimulating the production of reactive oxygen species and free oxygen radicals. Under normal

circumstances these molecules are rapidly removed from the system by free radical scavengers, like superoxide dismutase. These scavengers are however only present at low levels or are totally absent in the immature fetal brain tissue, resulting in oxidative stress and consequently neural apoptosis (Bonthius et al., 2003). Alcohol-induced oxidative stress also results in mitochondrial dysfunction. Mitochondria are the primary source of calcium storage and regulation, essential for neural viability and function. Under oxidative stress, mitochondria undergo mitochondrial permeability transition, a process where the mitochondrial membrane becomes perforated and releases it contents, which include calcium and cytochrome c. Both can activate caspases, which in turn induce both intrinsic and extrinsic apoptotic pathways (Bronner-Fraser, 2004).

- **Calcium imbalance**

The cellular calcium imbalance in the developing brain triggered by alcohol effectively integrates with all the above-mentioned modes of apoptotic induction. The apoptotic cascades initiated by alcohol's NMDA antagonist and GABA-mimetic properties are linked to a calcium influx by means of voltage-gated calcium channels (Burgoyne et al., 2004). Improper intracellular calcium accumulation consistently results in the generation of ROS, which is detrimental to neural cells (Bonthius et al., 2003). Intracellular accumulation of calcium also results in mitochondrial damage, which is discussed above. Furthermore, the anti-apoptotic factors contained in the mitochondrial membrane (*BCL-2* and *BCL-X*) lose their protective effect with the increase of intracellular calcium (Grisel et al., 2002). It is therefore necessary to also investigate pathways involved in the regulation of calcium levels in the brain. The phosphoinositide cascade is one such pathway, which is essential in modulating intracellular calcium levels. Studies have shown that at least one of the intermediates in this cascade, protein kinase C, is targeted by alcohol, and results in the inhibition of neural proliferation (Dick et al., 2002; Xu et al., 2003).

1.4.2 Cellular interaction

It was recently discovered that cell adhesion molecules (CAM) play an integral role in fetal brain development by facilitating neural migration, guiding cell-cell and cell-matrix adhesions, and facilitating signalling in the fetal brain (Heath and Nelson, 2002). The array of CAMs includes cadherins, selectins, integrins, mucins, and members of the IgG superfamily (Wilkemeyer et al., 2003). Studies have shown that mutations in the gene coding for one of the proteins from the latter group, neural cell adhesion molecule L1, lead

to disorders with similar features as FASD. It was consequently revealed that alcohol interferes with L1-mediated cell adhesion, which possibly contributes to the neural developmental disorders, and even the facial dysmorphology observed in FASD (Che and Chen, 2004; Lee et al., 2004). This interaction of alcohol with CAM is one example where alcohol does not result in apoptosis of the neural cells, but rather dysfunction of a specific molecule.

To date, no FASD linkage- or whole-genome association studies have been performed and no region of the genome has conclusively been linked to risk of FASD development. Furthermore, since the molecular basis of FASD is poorly understood and laboratory experimentation is costly, complex and ethically sensitive, there is strong motivation for using computational tools to extract data from existing knowledge (contained in the literature-, ontology- and genome databases) to identify possible candidate genes for FASD development, as the study approach.

1.5 AIMS AND OUTLINE OF STUDY

The principal objective of this project was to develop an innovative approach to identify genes that could potentially impact on disease development. This method should not be limited by the absence of a linked genetic locus, or known genetic causes. Furthermore, the system should be flexible regarding the number and variety of data sources that could be utilized.

In light of the current burden of FASD in many resource-poor communities in South Africa and the inconclusive search for susceptibility genes, computational identification offers an efficient approach to the identification of disease-causing genes. Although the approach described here was developed and modelled focusing on FASD, the objective was to develop an approach that would have applicability to any complex disease. This approach would be a cost-effective alternative to experimental methods used to identify disease genes, such as linkage analysis and genome-wide association studies.

The specific aims of this study were to:

1. Identify putative candidate genes for FASD, using text-mining and ontology database-linked expression data in the public domain.
2. Compile a prioritized sub-list appropriate for experimental analysis, from the original list of genes obtained through Aim 1.

3. Evaluate the appropriateness and validity of the prioritization approach.

Once candidate genes are selected, the next step was to evaluate the genetic variation and the potential functional impact of the variants within the genes. Ultimately a genetic association study would statistically confirm whether genetic variation within the candidate genes is possibly associated with FASD development. Therefore the second segment of this project focused on the evaluation of genetic variation within the regulatory regions of the top-ranked candidate gene, for inclusion in future association study analyses.

The specific aims of this section of the study were to:

1. Sequence the upstream untranscribed region and 5' untranslated regions of the top-ranked candidate gene in a randomly selected group of individuals from the Upington and De Aar study populations.
2. Analyse sequenced regions to identify genetic variation unique to these population groups.
3. Select a subset of SNPs within the candidate gene from all known SNPs, based on *in silico* evaluation of the putative impact on protein function and gene expression.

Chapter 2

Computational Selection and Prioritization of Genetic Targets for Fetal Alcohol Spectrum Disorders

2.1 INTRODUCTION

In this era of rapid technological advance and the promise of the $1000 genome, it is disappointing that not more is already known about the genetics of complex disease. The reason for this gap in knowledge seems to be due to the particular research processes applied to elucidating the genetics underlying these disorders.

As detailed in *Section 1.1*, two main approaches are used in investigating the genetics underlying disease – linkage analysis and association studies (Barnetche et al., 2005; Botstein and Risch 2003). Linkage analysis has a relatively low power to detect markers exhibiting a moderate influence on disease susceptibility (as is most often the case for complex disease genes) and therefore association studies are usually favoured to identify the genetic variants influencing complex disease risk. Whole-genome association studies eliminate the need for selecting candidate genes, and recently these investigations have started yielding notable success (Dempfle et al., 2006; Klein et al., 2005; Namkung et al., 2005; The Wellcome trust case control consortium, 2007). Although genotyping technologies have improved considerably and costs have been lowered, the platforms needed to perform whole-genome association studies are not readily available in many countries and are at the moment still relatively expensive. Candidate gene association studies therefore remain the most practical and frequently employed approach in disease gene investigation for complex disorders, necessitating efficient candidate selection (Daly and Day 2001).

To date, no FASD family linkage studies or genome wide association studies have been performed. Linkage studies require large family samples and this poses a significant challenge. Countries with the highest FASD rates are mostly resource-poor, possibly contributing to the reason why such studies have not yet been performed. Few candidate gene association studies investigating the effect of specific genetic polymorphisms on the risk of FASD development have been published. These studies have generally focused on the alcohol dehydrogenase enzymes gene family members and conflicting results have been obtained (see *Section 1.4*). Many other genes are likely to contribute towards the development of FASD and further investigation is required. In light of the current burden of FASD in many resource-poor communities in South Africa and the inconclusive search for susceptibility genes, computational identification offers a novel and efficient approach to the identification of disease-causing genes.

Recently, many computational candidate gene selection and -prioritization methods have been developed (Adie et al., 2006; Aerts et al., 2006; Franke et al., 2006; Freudenberg and

Propping 2002; George et al., 2006; Kent et al., 2005; Lopez-Bigas and Ouzounis 2004; Perez-Iratxeta et al., 2005; Tiffin et al., 2005; Turner et al., 2003; van Driel et al., 2005). These tools aim to identify and prioritize putative disease genes by modelling specific characteristics of known disease genes, or by focusing on known disease features (such as gene expression profiles or phenotype). However, there is a vast quantity of information and data sources available currently, and it is expected that a tool with the flexibility to include a large array of data sources would positively aid disease gene discovery.

The freely accessible tool Endeavour offers such an application (Aerts et al., 2006). This tool is based on the premise of ranking unknown candidate genes according to their similarity with a known set of training genes. In the absence of a linked genetic region (which is the case with FASD), all genes in the genome must be included as a starting point for candidate gene selection, which is not feasible when using this approach. CFG (convergent functional genomics) is an approach used to identify and prioritize candidate genes which relies on the cross-matching of animal model gene expression data with human genetic linkage data, as well as human expression- and functional data (Bertsch et al., 2005; Bertsch et al., 2005; Rodd et al., 2007). This approach has many parallels to the approach described here, as it prescribes a Bayesian-like methodology of reducing uncertainty through the combination of multiple independent lines of evidence, each by itself lacking sufficient power to confirm that a gene is a putative candidate gene, to produce a short list of high probability candidate genes (Rodd et al., 2007).

The approach of CFG relies principally on two lines of evidence – animal model data and human genetic linkage data. The approach developed here has the added advantage of allowing the inclusion of additional lines of evidence in the presence of limited expression studies in an animal model and the absence of linkage studies for FASD. Initially, the candidate gene selection method described by Tiffin et al. (2005) was employed (see *Section 1.2.3* for a description of the methodology), but in the absence of a candidate genetic region, this method resulted in a large candidate gene list, as it relies on the selection of candidate disease genes only according to their expression profiles. Therefore a new prioritization method was devised to rank genes from the candidate gene list for empirical investigation. In this process, a variety of relevant database sources are mined for candidate genes that exhibit characteristics relevant to disease phenotype. Genes were prioritized based on binary evaluation, where genes were assessed using criteria pertinent to FASD to mine various database sources and to create criteria-specific gene lists. For this approach an extensive selection of data sources (detailed in *Table 2.1*) was included for prioritization, in order to

comprehensively evaluate the candidate genes. However, this approach has the added flexibility of including additional data sources, if needed. Furthermore, the process is not a filtering process, where one expects unlikely candidates to be discarded, but rather prioritizes the genes within the list without any exclusion. This enables the researcher to obtain a sense of which genes are most likely to be involved in the disease, and still inspect the submitted list as a whole.

The prioritization method described here is based on a simulation of one possible way in which the researcher would usually select candidate genes i.e. by filtering through various data sources and selecting genes that exhibit the biological characteristics expected to indicate a link to the particular disease. This process could be laborious and not always intuitive, which implies that automated mining of data sources could be beneficial in knowledge discovery.

2.2 METHODS

2.2.1 Literature search

Abstracts related to FASD were obtained from the PubMed scientific literature database. In order to obtain all relevant literature, PubMed's automatic term mapping search of the literature might not be sufficient and a more robust search option of using Medical Subject Heading (MeSH) terms was implemented. MeSH is a controlled vocabulary used to describe the subject of each journal article in MEDLINE – imposing uniformity and consistency (US National Library of Medicine - PubMed tutorial, 2007). MeSH terms are arranged hierarchically, with more specific terms arranged beneath broader terms. Using MeSH terms when searching PubMed implies that all equivalent synonyms or lexical variants in English will be included in the search (US National Library of Medicine - PubMed tutorial, 2007). To compensate for the possibility that a specific article was not labelled with the MeSH term of interest, the term was also searched with a text word tag. Terms that are qualified with the text word field tag will be searched in the following fields – title, abstract, MeSH headings and subheadings, other term fields, chemical names of substances, secondary source identifiers and personal name as subject. Automatic inclusion of more specific MeSH terms related to the search term is excluded when this option is used. Literature related to FASD was obtained using the following query: "(fetal alcohol syndrome [MH]) OR (fetal alcohol spectrum disorder* [tw])" Limits: only items with abstracts, English.

2.2.2 Literature mining

The online literature mining tool Dragon Disease Explorer (DDE) was used to extract eVOC ontology terms from the body of literature. DDE provides summarized information from a body of submitted PubMed abstracts about frequency of occurrence of ontology terms within the text. This assists biologists in uncovering possible functional associations between disease and gene expression site. Following the method of Tiffin et al. (2005), only eVOC anatomy terms were used to extract the initial candidate gene list. Cell type terms were used to populate criteria lists for the binary filtering approach. Terms extracted matching to the developmental- and pathology ontologies were uninformative in this case (terms such as pathology or adult were extracted) and it was deemed that populating criteria lists using these terms would not contribute positively to the selectivity of the binary evaluation system. Therefore these terms were not further included.

2.2.3 Candidate gene selection

The method previously described by Tiffin et al. (2005) was used to extract candidate genes based on the information obtained from the literature mining. Figure 2.1 illustrates the process of literature- and data-mining used to select candidate genes. Briefly, this method ranks the extracted eVOC terms by calculating a ranking score for each associated eVOC term, according to the frequency of association and the frequency of annotation of the eVOC term. The four top-scoring eVOC terms were selected from the ranked list, and compared with eVOC terms annotated to genes within the Ensembl database v.33 (September 2005) to select candidates (Flicek et al., 2008).

Symbols and IDs assigned by the HUGO gene nomenclature committee (HGNC) are the identifiers preferably used here, as the HGNC is the global authority to assign unique and standardized gene symbols to human genes. The primary identifier for each gene annotated by the HGNC is the current approved gene symbol which is an acronym or abbreviation of the associated gene name. In addition to this, each entry is assigned a unique HGNC ID, which enables easy data tracking regardless of updates in the nomenclature of any given entry (Bruford et al., 2007).

The system allows for one mismatch, such that candidates selected are those annotated with at least three of the four top-scoring eVOC terms. This approach was tested by the authors on a subset of genes representative of those that might be selected by a linkage analysis study, and not the full complement of genes in the Ensembl database, as in the

current study. The original Python scripts were therefore altered for this application (see *Addendum A*).

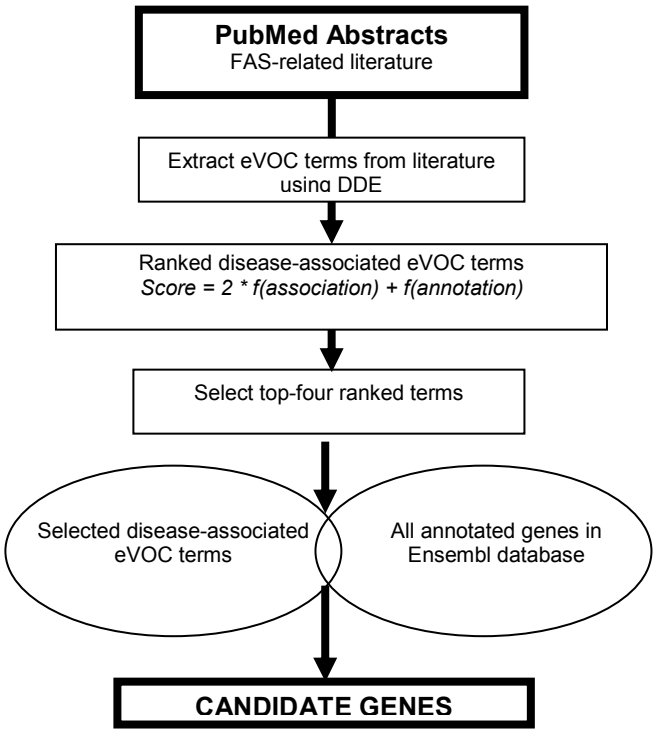

Figure 2.1: The method of integrated literature- and data mining to identify an initial list of putative candidate genes

2.2.4 Binary filtering and prioritization of candidate genes

The integrated literature- and data-mining approach to identify candidate genes focuses exclusively on anatomical sites related to the disease of interest, and results in a large list of genes. In order to obtain a more focused assessment of the most likely candidates from this gene list, other criteria pertinent to FASD were investigated.

Five main categories of criteria were used – cell type, biological process, homology, imprinted genes and phenotype simile. For each category there are multiple gene lists, each specified by one criterion (Table 2.1). A Python script was used to compare the criteria-specific gene lists generated to the candidate gene list (obtained from the integrated data- and literature-mining approach described above) in order to compile a binary matrix (see *Addendum A*).

Table 2.1: Summary of the criteria used to extract gene lists to compare to the master gene list, to create a binary grid. A total of 29 criteria were used to populate 29 gene lists.

CRITERIA CATEGORIES				
Cell type	**Biological Process**	**Animal model homology**	**Phenotype simile**	**Imprinted genes**
Glial cell Neuron Fibroblast Neuroepithelium	Apoptosis Development Brain Development Transport Signal Transduction	*Phenotype* Growth Behaviour/Neurological Craniofacial Nervous system related Embryogenesis *Timing* Pre-Embryonic Embryonic Fetal *Anatomy* TS[1]8-9 Ectoderm TS10-13 Neural Ectoderm TS14-26 CNS TS12-26 Head TS20-26 Cranium TS28 CNS	Mental Retardation Microcephaly Craniofacial Hyperactivity Growth Retardation	All known human imprinted genes

[1]TS – Theiller stage: A term used to denote the stage of development of a mouse as described by Theiler in "The House Mouse: Atlas of Mouse Development" (Springer-Verlag, New York, 1989)

The binary evaluation was performed as follows (illustrated in Figure 2.2): A gene in the candidate gene list was assigned a 1 when that gene was also present in the gene list obtained by a specific criterion. If the gene was absent from that list it was assigned a 0. For each of the genes the final binary score was calculated simply by summing all binary scores for each of the criteria used. All genes were then ranked based on this score, with those having higher scores being higher in the rank list. Genes in the candidate list that were present in most criteria lists (i.e. those genes obtaining the most 1-scores in the binary matrix) received the highest rank as candidates. This follows the premise that genes most commonly selected from additional independent sources possess characteristics that make them more promising candidates. Similarly, genes that were selected by only one or none of the additional criteria have a lower rank and are considered to be weak candidates.

A description of each category of criterion and the information used to assess the criteria are given below:

Candidate gene list	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	...
X gene 1	0	0	0	0	0	
X gene 2	1	1	1	1	0	
X gene 3	0	0	0	1	1	
X gene 4	1	0	0	0	0	
...						

BINARY FILTERING AND PRIORITIZATION

Candidate gene list	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	...
X gene 2	1	1	1	1	0	
X gene 3	0	0	0	1	1	
X gene 4	1	0	0	0	0	
X gene 1	0	0	0	0	0	
...						

Figure 2.2: Illustration of the binary filtering and prioritization process. Genes on the X chromosome (candidate gene list) were compared to the generated criteria lists. A gene in the candidate gene list was assigned a 1 when that gene was also present in the criteria gene list that it was evaluated against. If the gene was absent it was assigned a 0. For each of the genes on the X chromosome a binary score was calculated simply by summing all binary scores for each of the criteria lists used.

Cell Type

DDE was used to extract all eVOC cell type terms from the disease-related literature. Cell type ontology terms found to be associated with FASD were compared with eVOC terms annotated to genes within the Ensembl database to select a list of genes.

Biological Process

Disease-related literature contains terms describing functional aspects related to the disease. Dragon TF Association Miner (DTFAM) is an online tool for text-mining of PubMed abstracts to discover potential functional association of GO terms and diseases (Pan et al., 2004). DTFAM was used to extract all GO terms from the abstracts of disease-related literature. Of the terms extracted, terms falling in the molecular function (binding) and cellular component (membrane, nucleus, chromosome and intracellular) ontologies were not included in the analysis, as these terms were considered non-specific with regard to FASD and non-specific in general. Terms from the biological process ontology considered uninformative were also eliminated. This includes terms such as pathogenesis or lactation that would appear in the relevant literature due to subject matter described, and not because of relevance to disease. Genes annotated with the selected GO terms extracted from the literature were obtained from the Ensembl database v.33 (September 2005), and each individually used to populate a criteria list.

Animal model homology

Animal models offer major contributions to the understanding of human disease. Although many different animal models for FASD have been developed (Cudd, 2005), the mouse model seems to correlate best to the effects of prenatal alcohol exposure observed in humans (Sulik, 2005). MGD documents the mouse as a model system for human biology and disease process research (Eppig et al., 2005). MGD integrates genetic and genomic data for the mouse, including sequence sets, mapping details, GO annotations, allele descriptions and mutant phenotype characteristics. Furthermore MGD provides a curated set of mammalian orthologues (Blake et al., 2006).

Human orthologues to the following categories of mouse genes were selected:

- Genes associated with phenotypes affected by prenatal alcohol exposure
- Genes expressed at different developmental stages
- Genes expressed in the developing brain

Phenotype simile

It is assumed that similar phenotypes may be influenced by similar genotypes (Oti and Brunner 2007). The main characteristics of FASD are growth retardation, distinct craniofacial dysmorphology and CNS dysfunction. The neurodevelopmental consequences of CNS dysfunction due to prenatal alcohol exposure include cognitive deficits (often mental retardation), executive functioning deficits, motor functioning delays and problems with attention, hyperactivity and social skills (Welch-Carre, 2005). Terms describing key phenotypes associated with FASD (mental retardation, microcephaly, craniofacial, hyperactivity and growth retardation) were used to search for genes in the GeneCards catalogue v2.36 (April 2007) (Safran et al., 2003). Genes linked to these phenotype terms were used to create the criteria lists.

Imprinted Genes

Genomic imprinting is an epigenetic mechanism, resulting in the preferential expression of either the paternal or maternal allele of certain genes (Surani, 1998). One of the characteristics of imprinted genes are differential allele-specific DNA methylation, which is usually localized to regions known as differentially methylated regions (DMRs) of which there are generally two varieties. Firstly, there are DMRs that acquire differential methylation during somatic development, in a tissue-specific manner (Reik and Walter, 2001). The second variety of DMRs are referred to as imprinting control regions (ICRs), which are differentially methylated in all tissues, throughout development. ICRs are the

principal regulators of gene expression within imprinted domains, often over large distances (Robertson et al., 2006).

The expression of many prokaryotic and eukaryotic genes is regulated through the methylation of DNA (Lim and van Oudenaarden 2007). Animal studies have shown that in utero ethanol exposure inhibits fetal DNA methylation (Garro et al., 1991; Garro et al., 1991; Garro et al., 1991; Valles et al., 1997). Since DNA methylation and imprinting play an important role in the regulation of gene expression during embryogenesis (Garro et al., 1991; Garro et al., 1991; Wagschal and Feil 2006) and consequent development, ethanol-associated alterations in fetal DNA methylation may contribute to the developmental abnormalities seen in FASD. One of the criteria gene lists therefore contained all known imprinted genes, obtained from the imprinted gene catalogue (Glaser et al., 2006) and the genomic imprinting website (Jirtle et al., 2000).

As a way of further assessing the list of 87 prioritized FASD candidate genes they were cross-matched against candidate genes for alcoholism, obtained using CFG (Rodd et al., 2007). Although the two phenotypes are very different, one would expect some overlap in prioritized candidate genes since many of the mothers of FASD children suffer from alcoholism.

2.2.5 Evaluation of biological significance of prioritized genes

Protein-protein interactions, functional enrichment and common promoter element binding sites were investigated for the top-ranked genes (i.e. those with the highest binary score) to assess their biological significance as candidates for FASD. In comparison, the lowest-ranked genes were similarly evaluated to assess the validity of the ranking system in selecting biologically relevant genes from the original candidate gene list.

- **Protein-protein interactions**

Understanding interactions between proteins involved in common cellular functions may indicate how such interactions can influence disease outcome. Protein-protein interactions were analysed using data contained in the STRING (Search tool for the retrieval of interacting genes/proteins) database v7.1 (Von Mering et al., 2007). The STRING database provides a comprehensive source of protein-protein association evidence under a common framework. STRING integrates protein-protein interaction data from both experimental

evidence databases – such as BIND, DIP and MINT – as well as inferred protein-protein interactions obtained by using de novo prediction tools (such as Predictome), or functional grouping databases (such as Reactome or KEGG). The user can select which lines of evidence to use, and each predicted association in the database is assigned a confidence score, based on comparison to a common reference set of true associations. The top-ranked candidate genes and randomly selected low-ranked genes were used as input, and both predicted and known protein-protein interactions based evidence were selected as evidence. A medium confidence score for evidence was selected (50%).

- **Functional enrichment analysis using DAVID**

DAVID (Database for annotation, visualization and integrated discovery) is an online tool that integrates genomic functional annotations to reveal biologically relevant enrichment in a gene list (Dennis et al., 2003). DAVID promotes functional discovery through exploration of biochemical pathway maps, functional classification using GO terms and conserved protein domain architecture. Data from various sources are integrated into DAVID, including GenBank, UniGene, RefSeq, Locuslink, KEGG, OMIM and GO. The top-ranked genes were submitted as a list to DAVID 2007 (January 2007), which was then compared to a background gene list to assess functional enrichment within the list. The background list can either be all genes in the human genome, or a sub-set of genes. Two analyses were performed – firstly with the original candidate gene list of 10174 genes as background, and secondly using the *Homo sapiens* default background list from the DAVID website as background.

2.2.6 Promoter element binding site analysis

To investigate potential drivers of transcription initiation of the top-ranked candidate genes and associate the prioritized genes better to the FASD phenotype, mammalian TFBS were predicted. This was done using matrix models in Transfac database v9.4 for the promoters of all prioritized genes. Thresholds that correspond to the minimum number of false positive predictions as defined by minFP profiles in Transfac were used. The same process was applied to 10255 human promoters according to Bajic et al. (2006). Using the methodology of contrasting target promoter set with the background set of 10255 human promoters (Bajic et al., 2004), the most dominant promoter elements were determined. A promoter element is defined as a TFBS and the strand where it is predicted, or as a pair of these if they are at the maximum distance of 50 nucleotides. This analysis was performed by Prof.

Bajic and Dr. Hofmann at the South African National Bioinformatics Institute, University of the Western Cape.

2.3 RESULTS

2.3.1 Integrated literature- and data mining for candidate gene selection

According to the method described by Tiffin et al. (2005), DDE was used to extract eVOC anatomical terms from the body of literature, whereafter they were used to extract candidate genes from the Ensembl database. This method extracted a list of 10174 genes, a reduction of 70.3 % from the original 34294 genes in the Ensembl database.

2.3.2 Binary filtering and prioritization of candidate genes

In order to select the most likely candidates from the initial candidate gene list, these genes were ranked according to the number of additional criteria they matched (Table 2.1). The top-ranked genes (in ranked order) are shown in Table 2.2. Fibroblast growth factor receptor 1 (*FGFR1*) was the top-ranked gene, present in 17 of the 29 criteria gene lists, followed by Msh homeobox homolog 1 (*MSX1*), present in 16 of the 29 criteria lists. Fibroblast growth factor receptor 2 (*FGFR2*), Forkhead box G1B (*FOXG1B*) and Homeobox A1 (*HOXA1*) were present in 15 of the 29 criteria lists, followed by a group of 4, 17, 14 and 47 genes present in 14, 13, 12 and 11 criteria lists, respectively. This group of 87 genes was used as the prioritized candidate gene list for further analyses (see *Addendum B*). This cut off (present in 11 of the 29 criteria lists) is an arbitrary cut-off used to select an appropriately sized group of prioritized genes to investigate, a number which could typically also be suitable for a candidate gene association study.

Genes from the candidate gene lists that matched one or none of the criteria were considered to be unlikely candidates. Based on this premise, these 5055 genes (50%) from the candidate gene list were ranked as weak candidates. A negative control set were generated to assess the validity of the ranking method, by randomly selecting 87 genes of the subset matching to no criteria (see *Addendum A* for Python script).

Table 2.1: Selected top-ranked candidate genes for FASD identified using binary matrix filtering

Criteria matched	HGNC ID	Location	Function
17/29	*FGFR1*[1]	8p11.2	Involved in limb induction, plays a role in bone elongation modulation
16/29	*MSX1*[2]	4p16.3-p16.1	Potential repressor function in cell cycle progression, transcription repressor
15/29	*FGFR2*[1]	10q26	Involved in vertebral development, regulator of bone formation and osteoblast activity
15/29	*FOXG1B*	14q13	Embryonic transcriptional regulator critical in brain development
15/29	*HOXA1*	7p15.3	Placement of hindbrain segments along the anterior-posterior axis
14/29	*BMP4* [2,3]	14q22-q23	Regulating myogenesis, induces apoptosis and chondrogenesis
14/29	*FGFR3*[1]	4p16.3	Negative regulator of bone growth promotion, inhibition of chondrocyte proliferation and differentiation
14/29	*GNAS*[2]	20q13.2-q13.3	Involved as modulator in various transmembrane signalling systems mediating the effect of parathyroid hormone
14/29	*PAX6*	11p13	Key regulator of eye, pancreas and CNS development and regulates glial precursors in ventral neural tube

[1]Members of/linked to the MAPK signalling pathway; [2]Members of/linked to the TGF-β signalling pathway; [3]Members of/linked to the Hedgehog signalling pathway.

2.3.3 Evaluation of biological significance of prioritized genes

- **Protein-protein interactions**

The list of top-ranked candidate genes (top-ranked 87 genes), and unlikely candidates (randomly selected low-ranking 87 genes) were submitted to the STRING database to assess known protein-protein interactions. Figure 2.3 (a) shows the STRING network of interactions for the top-ranked genes. The lines represent an interaction, with the line thickness indicating the confidence level for the interaction evidence – i.e. thick lines indicate high confidence and thin lines low confidence. Significantly fewer interactions were present in the low-ranked list (Figure 2.3 (b)). The gene products found to interact with evidence with high confidence levels (≥ 90%) are summarized in *Addendum B*.

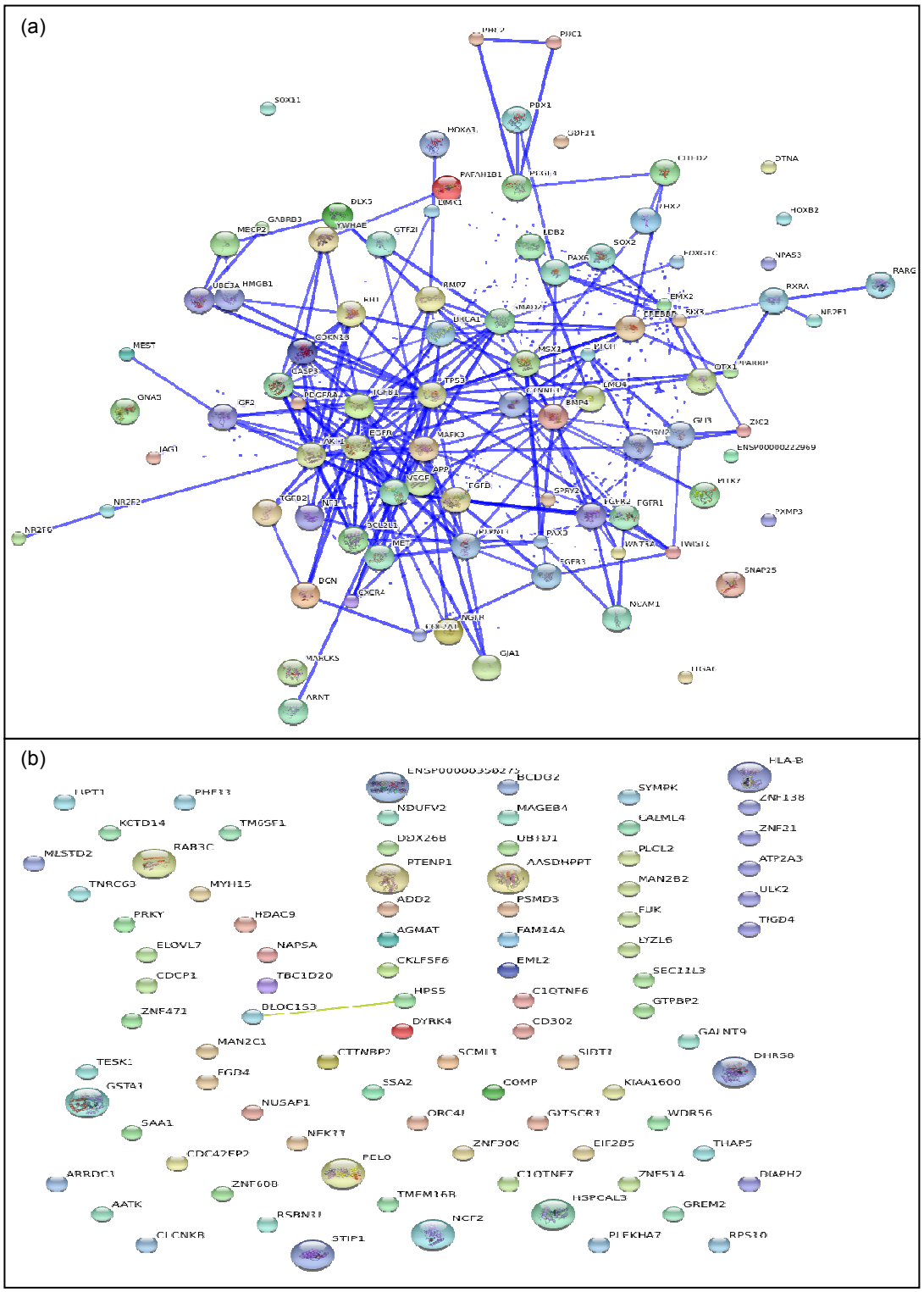

Figure 2.3: The STRING network of known protein-protein interactions among the (a) 87 top-ranked candidate genes for FASD and (b) 87 randomly selected bottom-ranked genes. The network edges represent the predicted functional associations. The network edges represent the predicted functional associations, with the thickness of the line representing the confidence of the predicted interaction.

- **Functional enrichment analysis using DAVID**

DAVID (Dennis et al., 2003) was used to assess functional enrichment within the top-ranked candidate gene list. Firstly, the analysis focusing on pathway maps highlighted a number of pathways significantly represented within the gene list, with the transforming growth factor-β (TGF-β) signalling pathway being most over-represented within the list (Table 2.3). This enrichment was not observed on the low-ranked gene list. Furthermore, significant enrichment of GO terms was observed in the top-ranked list for the GO categories related to function (biological process and molecular function). The GO terms found to be significantly enriched for the top-ranked gene list are shown in *Addendum B*.

Table 2.3: Biological pathways significantly over-represented among the top-ranked candidate genes. The gene count indicates how many genes from a particular pathway were present in the candidate gene list of 87 genes. Note that varying *P*-values were obtained depending on the background list used

Pathway	Gene Count	*P*-value[1]	*P*-value[2]
TGF-b signalling pathway	9	6.7×10^{-6}	1.0×10^{-6}
Hedgehog signalling pathway	7	3.8×10^{-5}	3.6×10^{-5}
MAPK signalling pathway	13	7.8×10^{-5}	6.0×10^{-4}
Adherens junction	7	3.5×10^{-4}	5.1×10^{-4}
Cell cycle	8	3.6×10^{-4}	0.001
Neurodegenerative disorders	5	7.5×10^{-4}	0.002
Regulation of actin cytoskeletan	9	0.004	0.014
Focal adhesion	9	0.004	0.022
Gap junction	6	0.006	0.008
Cytokine-cytokine receptor interaction	9	0.011	0.002
Epithelial cell signalling in H.Pylori infection	4	0.018	0.029

[1]P-value obtained using the Homo sapiens gene list as a background list to the top-ranked candidate genes; [2]P-value obtained using the original candidate gene list as a background list to the top-ranked candidate genes

- **Promoter element binding site analysis**

As shown in Tables 2.4 and 2.5, the promoter analysis detected 15 transcription factors (TF) that appear in promoter elements (PE) or pairs of PE that are significantly statistically enriched in the target promoter set as opposed to the background set. The conditions for the selection of PE for Tables 2.4 and 2.5 were that PE (or their combination) has to appear in at least 5% of promoters in the target set and to have over-representation index (ORI) (see Bajic et al., 2004) of at least 2. These are AP-2, C/EBP, E2F, ETF, LEF1, MAZ, MAZR, MZF1, Pax-4, Sp1, Spz1, TATA, TFII-I, VDR, ZF5. In Tables 2.5 and 2.6, PE or their combinations that have been found in significantly enriched proportions relative to the

background promoter set, are denoted by a + sign in the column of the over-representation index (ORI). Further analysis suggests that TF that potentially bind these transcription factor binding sites (TFBS), are part of the group of TF that are dominant transcriptional regulators of our promoter target set (Tables 2.4 and 2.5). Additional results from the promoter element binding site analysis are shown in *Addendum B*.

Table 2.4: Promoter elements found to be enriched in the target promoter set relative to the background promoter set. The criteria for selecting PE as enriched was that it has to appear in at least 5% of the target promoter sequences and to have over-representation index (ORI) of at least 2. PE that appear in statistically significant proportion in the target set are denoted by + in the ORI column.

Promoter elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
-1 MAZR	11.5685	5.7	2.07	$9.0x10^{-5}$	$2.0x10^{-5}$	31	212	544	10255	0.002
+1 MAZR	5.6322	5.15	2.08	$5.0x10^{-5}$	$2.0x10^{-5}$	28	213	544	10255	0.033
-1 TATA	2.9231	16.36	9.76	$1.7x10^{-4}$	$1.0x10^{-4}$	89	1001	544	10255	0.002
-1 TFII-I	2.8865	18.2	10	$1.7x10^{-4}$	$1.1x10^{-4}$	99	1025	544	10255	<0.001
-1 MAZ	2.6342	29.96	20.03	$4.0x10^{-4}$	$2.2x10^{-4}$	163	2054	544	10255	<0.001

Table 2.5: Pairs of promoter elements found to be enriched in the target promoter set relative to the background promoter set. The criteria for selecting pairs of PE as enriched was that it has to appear in at least 5% of the target promoter sequences and to have over-representation index (ORI) of at least 2. PE that appear in statistically significant proportion in the target set are denoted by + in the ORI column.

Pairs of promoter elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
-1 MZF1/+1 E2F	17.10	6.62	1.47	0.0001	<0.0001	36	151	544	10255	<0.001
-1 LEF1/-1 Pax-4	14.67	6.62	1.77	0.0001	<0.0001	36	182	544	10255	<0.001
-1 C/EBP/+1 VDR	10.15	6.43	1.85	0.0001	<0.0001	35	190	544	10255	<0.001
+1 C/EBP/+1 VDR	9.10	6.99	2.42	0.0001	<0.0001	38	248	544	10255	0.042
-1 MAZ/-1 VDR	9.02	9.19	2.89	0.0002	0.0001	50	296	544	10255	<0.001
-1 MZF1/-1 MZF1	8.46	5.51	1.6	0.0001	<0.0001	30	164	544	10255	0.040
-1 ETF/-1 VDR	7.67	7.54	2.68	0.0001	<0.0001	41	275	544	10255	0.024
-1 AP-2/-1 ETF	6.89	12.87	5.11	0.0002	0.0001	70	524	544	10255	<0.001
-1 MAZ/+1 Sp1	6.45	9.56	2.63	0.0001	0.0001	52	270	544	10255	<0.001
-1 Spz1/-1 Spz1	6.05	14.34	5.57	0.0002	0.0001	78	571	544	10255	<0.001
-1 VDR/-1 Spz1	5.43	14.34	6.54	0.0002	0.0001	78	671	544	10255	<0.001
-1 ETF/-1 E2F	5.40	11.76	5.22	0.0002	0.0001	64	535	544	10255	0.007
-1 Spz1/-1 VDR	5.25	13.6	6.58	0.0003	0.0001	74	675	544	10255	0.013
-1 VDR/+1 ZF5	4.43	14.15	6.59	0.0002	0.0001	77	676	544	10255	0.001

2.4 DISCUSSION

This study was aimed at selecting a highly likely group of candidate genes for susceptibility to FASD, in the absence of genetic linkage evidence. A computational approach to candidate disease gene identification is proposed as an effective first line of candidate gene identification for a complex disease such as FASD. Mining of gene expression data was used to generate an extensive list of candidate genes which were compared to filtered criteria specific gene lists using 29 criteria to select the most likely candidate genes. The prioritization method described here is based on a computational model of a researcher's approach to selecting candidate genes, i.e. based on published information; but may also select non-intuitive candidate genes. In summary, various relevant database sources are accessed to establish whether a candidate gene and its product exhibit the biological characteristics consistent with that particular disease.

2.4.1 Candidate gene selection and -prioritization

A method that employs an integrative literature- and data mining approach to select candidate genes was used to select candidate genes for FASD (Tiffin et al., 2005). This method extracted a gene list of 10174 genes. This list is relatively unspecific, and is likely to have a high false-positive rate. The most plausible explanation for the selection of such a large, ambiguous list is a lack of detailed information about the source of cDNA libraries, with the result that more general terms from higher up the ontology hierarchy are often used for annotation of the gene. This led to the formulation of a prioritization method to rank genes from the candidate gene list using many different data sources for laboratory investigation of individual candidate genes. A binary evaluation method was used to rank the candidate genes in the list, facilitating the selection of 87 top-ranked genes as the most likely candidate genes for further investigation.

Further analysis with available online tools such as DAVID and STRING highlighted protein-protein interaction, functional enrichment and probable biological significance among the top-ranked genes. STRING was used to investigate protein-protein interaction among the prioritized candidate genes, and highlighted a group of genes that interact (Figure 2.3). The candidate gene selection method described here focuses on gene annotation, and it is therefore possible that the top ranking genes are better annotated than low-ranked genes. Consequently, a potential source of bias is that disease-causing proteins may display a higher degree of interconnectivity, simply because they are better studied. The absence of protein-protein interaction among the low-ranked genes is

therefore not necessarily a reflection on the level of interaction but may be related to the level of understanding of the gene and its function. However, it has also been shown that disease genes exhibit an increased propensity for their protein products to interact with one another, tend to be coexpressed in specific tissues, and display consistent functions with respect to all three GO categories (Goh et al., 2007; Ideker and Sharan, 2008). One can therefore expect that a protein-protein network analysis (as well as a functional enrichment approach such as DAVID) would give some indication of a group of genes' likelihood of being putative disease genes. It is accepted that the genes underlying complex disease (such as FASD) will be plentiful and the interactions between these factors are likely to be intricate. For this reason, STRING is a useful tool to highlight genes within the top-ranked gene list that interact and that may have a cooperative effect on disease outcome.

DAVID elucidates functional enrichment and biological significance within the top-ranked gene list, and highlighted the TGF-β and Mitogen-Activated Protein Kinase (MAPK) signalling pathways as primary candidate pathways for FASD development.

As a way of further assessing the list of 87 prioritized FASD candidate genes they were cross-matched against candidate genes for alcoholism, obtained using CFG (Rodd et al., 2007). Although the two phenotypes are very different, one would expect some overlap in prioritized candidate genes since many of the mothers of FASD children suffer from alcoholism. The two prioritized candidate gene lists (87 genes for FASD and 65 for alcoholism) had only two high priority candidate genes in common – GNAS complex locus (*GNAS*) and high mobility group protein B1 (*HMGB1*). The remaining 63 candidate genes for alcoholism were also present in the initially selected list of 10174 genes, but were ranked below the arbitrary cut-off of 11/29 criteria used to select the highly prioritized candidate list for FASD.

Incorporating the set of alcoholism genes as a selection criterion into the binary evaluation method only added two more genes to the prioritized list. These were G1/S-specific cyclin-D1 (*CCND1*) and insulin-like growth factor I receptor (*IGF1R*). Both gene products contribute to cell proliferation and differentiation, and exhibit characteristics that also make them likely candidate genes for FASD. However, neither directly interacts in the two main prioritized pathways (TGF-β or MAPK signalling pathway). This comparison shows that these two related diseases (due to the involvement of alcohol in both) have potentially common genetic factors, but that they also exhibit diversity in terms of genetic susceptibility. This gene list was therefore not included in the final binary filtering analysis.

2.4.2 Prioritized pathways – relevance to FASD development

- **TGF-β signalling pathway**

FASD encompass a range of complex syndromes, suggesting that the genetic factors underlying susceptibility to FASD may be plentiful and the interactions between these factors, as well as environmental factors are likely to be intricate. The computational approach described here highlights genes that are important players in various signalling pathways, in particular the TGF-β and MAPK pathways. These genes play pivotal roles during embryogenesis and development (Table 2.2) and have a potential role in the distinct characteristics associated with FASD, i.e. CNS dysfunction, craniofacial abnormalities and growth retardation. CNS dysfunction is the most severe and permanent consequence of in utero alcohol exposure. These observations make the TGF-β signalling pathway an interesting focus point, as it is essential in both fetal development and also CNS development (Gomes et al., 2005).

TGF-β signalling controls a diverse array of cellular processes, including cell proliferation and apoptosis, cell differentiation and specification of cellular phenotypes and developmental fate (Shi and Massague 2003). TGF-β is also important in neuronal migration and axonal growth, and regulates the formation of various craniofacial structures (Chai et al., 2003; Chai et al., 2003; Miller and Luo 2002).

Early exposure to ethanol inhibits such TGF-β regulated processes as cortical cell proliferation and neuronal migration, disrupts axonal growth and up-regulates cell adhesion molecule expression (Miller and Luo 2002). It can therefore be suggested that members of the TGF-β signalling pathway interact with ethanol, and/or its metabolic breakdown products, and that ethanol may have a detrimental effect on the efficiency of this developmentally essential pathway. Investigating the role of TGF-β components present among the top-ranked genes may clarify part of the genetic component contributing to susceptibility for FASD development.

The hypothesis that TGF-β signalling pathway genes may be involved in FASD susceptibility is even more compelling when considering the major role of this pathway in neuronal apoptosis. Several studies have shown that alcohol suppresses neuronal activity, resulting in a pro-apoptotic environment in the developing brain (Farber and Olney 2003; Farber and Olney 2003; Farber and Olney 2003; Ikonomidou et al., 2000; Thiery, 2003;

Thiery, 2003; Thiery, 2003). Alcohol-induced neural apoptosis has been observed throughout the developing CNS, including all levels of the spinal cord, brain stem, cerebellum, midbrain and forebrain. Furthermore, alcohol has been observed to diminish neurons from various parts of the developing visual-, auditory- and memory systems of the developing brain (Farber and Olney 2003). This pro-apoptotic effect of alcohol provides a probable explanation for the long-term CNS dysfunction and diminished brain size associated with FASD, and could be mediated by the TGF-β pathway. Alcohol has an array of molecular pathway targets and modes of inducing apoptosis and the candidate disease genes selected using this method have a strong role to play in apoptosis.

Genetic mutations in members of the TGF-β signal pathway generally result in tumorigenesis and have been repeatedly linked to human cancer (Garrigue-Antar et al., 1995; Garrigue-Antar et al., 1995; Garrigue-Antar et al., 1995; Garrigue-Antar et al., 1995; Hahn et al., 1996; Hahn et al., 1996; Hahn et al., 1996; Hahn et al., 1996; Jakowlew, 2006; Jakowlew, 2006; Jakowlew, 2006; Jakowlew, 2006; Markowitz et al., 1995; Markowitz et al., 1995; Markowitz et al., 1995). TGF-β dysfunction is also causal for hereditary hemorrhagic telangiectasia (McAllister et al., 1994), corneal dystrophy (Mashima et al., 2000), Camurati-Engelmann Disease of bone (Saito et al., 2001) glomerulonephritis (Isaka et al., 1996), scar formation (Shah et al., 1995), keloids (Lee et al., 1999), pulmonary fibrosis (Khalil and Greenberg 1991), and liver cirrhosis (Castilla et al., 1991). Recent studies also propose a role for TGF-β signalling in Alzheimer's disease pathology (Das and Golde 2006; Tesseur et al., 2006; Tesseur et al., 2006; Tesseur et al., 2006). However, no such link has to date been proposed between genetic susceptibility to FASD development and disruption of the TGF-β pathway. Given the above-mentioned experimental evidence, the TGF-β pathway, and specifically its components that were top-ranked using this computational approach, is an attractive focus for a genetic association study.

- **MAPK signalling pathway**

The MAPK pathway transmits a large variety of external signals, leading to a wide range of cellular responses, including growth, differentiation, inflammation and apoptosis (Krens et al., 2006). This pathway is very complex and includes many protein components. MAPK-pathway components have been shown to be involved in both the initiation and regulation of meiosis, mitosis, and post-mitotic functions, and in cell differentiation by phosphorylating a number of transcription factors (Orton et al., 2005).

The MAPK signalling pathway can be activated by a variety of stimuli, including growth factors, cytokines and differentiation factors (Krens et al., 2006) as well as external stress factors, such as alcohol (Aroor and Shukla 2004). Recent studies have investigated the effect of controlling second-messenger signalling on neuronal migration in a mouse model of FASD (Kumada et al., 2006). It was shown that experimental manipulation of these second-messenger pathways, through stimulating calcium- and cGMP signalling or inhibiting cAMP signalling, completely reversed the action of ethanol on neuronal migration in vitro as well as in vivo. Each investigated second messenger had multiple but distinct downstream targets, including MAPK.

- **Hedgehog signalling pathway**

The hedgehog signalling pathway also received a highly significant ranking among the pathways identified to be enriched within the candidate list. The hedgehog signalling pathway is a key regulator of embryonic development and is highly conserved. Knock-out mouse models lacking components of this pathway have been observed to develop malformations in the CNS, musculoskeletal system, gastrointestinal tract and lungs (Ingham and McMahon 2001).

FASD animal models portray a strikingly similar craniofacial phenotype to mouse models treated with antibodies that block Hedgehog signalling components, specifically the sonic hedgehog (Shh) molecule (Ahlgren and Bronner-Fraser 1999; Ahlgren and Bronner-Fraser 1999; Cartwright and Smith 1995; Cartwright and Smith 1995; Chen et al., 2000). Further studies to expose the role of Shh in fetal alcohol syndrome, showed that alcohol resulted in a significant decrease in Shh levels in the developing embryo, as well as a decrease in the level of other transcripts involved in Shh signalling. Furthermore it was observed that the addition of Shh after ethanol treatment led to fewer apoptotic cranial neural crest cells, resulting in a significant decrease in craniofacial anomalies (Ahlgren et al., 2002). These results give compelling support that the components of the Hedgehog signalling pathway may also be important in the genetics of FASD.

2.4.3 Transcriptional regulators of the prioritized genes

All TFBS that are found to be statistically significant for FASD are known to be involved in gene expression and regulation in the CNS, endocrine system or development. The AP-2 family of TF is crucial for neural gene expression and neuronal development (Damberg,

2005); C/EBP is involved in neuronal signalling (Calella et al., 2007); the E2F family of TF is one of the key controllers of cell-cycle and has a known role in pathways controlling neuron death (Greene et al., 2007); ETF, the epidermal growth factor receptor-specific TF, is implicated in neuroblastoma (Itoh et al., 1992); LEF1 is expressed in the nerve system of mammals (van Genderen et al., 1994); MAZ is involved in Hodgkin's disease and paraneoplastic cerebella dysfunction (Bataller et al., 2003) and during neuronal differentiation (Okamoto et al., 2002); MAZR is implicated in the development of mouse limb buds (Kobayashi et al., 2000); MZF1 is involved in development (Perrotti et al., 1995) and implicated in the control of the BACE1 gene related to Alzheimer's disease (Lange-Dohna et al., 2003); Pax-4 is involved in the endocrine system and development (Tayaramma et al., 2006); Sp1 has multiple roles, but, for example, controls expression of Na+,K+-ATPase in neuronal cells (Benfante et al., 2005); Spz1 is involved in cell-proliferation (Hsu et al., 2005); TATA binding proteins are implicated in various processes involved in brain (Riazi et al., 2005); the TFII-I transcription factor family is implicated in craniofacial development of humans and mice (Tassabehji et al., 2005); VDR is associated with increased risk of schizophrenia (Handoko et al., 2006); and ZF5 is implicated in neuroblastoma differentiation (Dimitroulakos et al., 1999). These results support the prioritization of biologically relevant candidate disease genes.

2.5 CONCLUSIONS

The results obtained in this study suggest that making a clinically-informed selection from the evidence obtained from literature- and database-mining is an effective approach for candidate disease gene selection and -prioritization. The main limitation of this approach is that it is primarily based on gene annotation, and that it is therefore biased towards selecting better annotated genes. Furthermore, some clinical understanding of the disease aetiology is needed to aid the clinically-informed binary evaluation, and this process could be partly subjective and researcher-specific. The effectiveness of this approach critically depends on the genes under investigation being clearly defined both molecularly and physiologically, in order to avoid erroneous associations. A multitude of biological processes are affected by the insult of alcohol exposure, particularly given a predisposing genetic background. FASD as a range of developmental disorder presents with a spectrum of structural, behavioural and neurocognitive disabilities, which complicates this process of clearly defining focus. This is evident when considering the ambiguous results obtained when using the method that only considers general anatomical terms to select candidate genes (Tiffin et al., 2005). This

encouraged the inclusion of the binary prioritization technique to further enhance the selection process. In an attempt to further focus the binary prioritization method, I considered weighting certain criteria that are most likely to provide useful information for gene prioritization. However in light of the limited knowledge regarding the molecular aetiology of FASD, and of complex disease in general, I decided that such an approach might be too subjective, and would not constructively add to the analysis.

A further limitation of employing this approach in selecting candidate genes for a developmental disorder lies in the limited knowledge available regarding the mechanisms involved in such a disorder. The developing organism undergoes many rounds of pattern formation, generating complexity with each ensuing round of cell division and with cell differentiation. Even though the pathways identified using this technique are general fundamental role players in embryogenesis and development, the technique allowed the focus to fall on specific candidate genes within these pathways for investigation.

This approach is less expensive in comparison to the expected cost of performing wet lab experimentation to identify candidate genes for FASD. However, the human interaction needed throughout this process increases the effort of performing the evaluation, and this level of involvedness complicates the full automation thereof. One of the main complications in the automation of this technique is that it is primarily based on data that is available in the public domain and it must be accepted that not all information available is correct or current. It is expected that automation of this approach will involve a web wrapper agent, to allow constantly updated data to be available for the evaluation process.

The computational approach described here has been used to select and refine a 'most likely' candidate gene list according to known characteristics of FASD. It was demonstrated that it is possible to identify likely candidates that are biologically relevant to the disease, and therefore appropriate for gene association studies. By refining the candidate gene list for FASD using a binary evaluation approach, a subset of biologically relevant candidate genes for experimental validation was selected.

.

Chapter 3

Validation of the computational candidate gene prioritization method – X-linked mental retardation

3.1 INTRODUCTION

3.1.1 Rationale

Many computational candidate gene selection and -prioritization methods have been described (Adie et al., 2006; Aerts et al., 2006; Franke et al., 2006; Freudenberg and Propping, 2002; Kent et al., 2005; Lopez-Bigas and Ouzounis, 2004; Perez-Iratxeta et al., 2005; Tiffin et al., 2005; Turner et al., 2003; van Driel et al., 2005) that aim to select and prioritize putative disease genes by modelling specific characteristics of known disease genes, or by focusing on known disease features, such as gene expression profiles or phenotype. *Section 1.2* provides an outline of some of the methods that are currently available and their mechanisms of candidate gene selection.

It is essential to assess the effectiveness of a computational candidate gene selection method in selecting appropriate candidates. The ultimate test of validity would certainly be to evaluate the gene-disease association by performing a laboratory-based association study. However, the validation method most often employed in published methods is to apply the method to diseases with known genesis to assess whether the method correctly selects the known disease gene(s). This approach could potentially over-estimate the accuracy of the computational approach, especially if the diseases that the method is tested on are not complex diseases. Multiple genes contribute to a complex disease, and usually there will not be a single gene that exerts a major effect on the development of such a disease. For this reason the computational method should select multiple candidate genes in a test scenario, which would not be the case for a single-gene disorder. Still, the approach of testing the method on a disease with known aetiology is a cost-effective and swift way of establishing an estimate of the computational approach's accuracy and appropriateness.

This approach was therefore used to assess the computational method described in Chapter Two. It was decided to use X-linked mental retardation (XLMR), a group of diseases with over-lapping phenotypes and multiple genes involved, as the "test disease". The following points were taken into consideration:

- XLMR is a set of heterogeneous disorders of which some of the underlying genetics is known. Similar to FASD, XLMR is an umbrella term for several syndromes where mental retardation is a main feature and linkage analysis has indicated a causative link to the X chromosome.

- In contrast to FASD, most forms of XLMR are due primarily to mutations in a single gene and therefore a key feature of XLMR is locus heterogeneity (See *Section 3.1.2* below for a further discussion of the heterogeneity of XLMR). Other examples of diseases where mutations in a single gene cause pathogenesis, and locus heterogeneity is also present, include retinitis pigmentosa (Barragan et al., 2008; Barragan et al., 2005; Tong et al., 2006), muscular dystrophy (Day et al., 1999; Fendri et al., 2006; Haravuori et al., 2004) and deafness (Goldstein and Lalwani, 2002; Lezirovitz et al., 2008; Yang et al., 2005). This layer of genetic complexity makes XLMR a more appropriate test disease than a single-locus disorder in this scenario.
- Many forms of XLMR exist where the genetic causes have not been established. Furthermore, it is plausible that X-linked risk factors that predispose to, but do not cause mental retardation, exist. The method could therefore be applied to evaluate the effectiveness of the approach in prioritizing known XLMR genes, and at the same time uncover unknown candidate genes and emerging pathways for XLMR.

3.1.2 X-linked mental retardation (XLMR)

The American Association on Mental Retardation (now the American Association on Intellectual and Developmental Disabilities) defines mental retardation as a disability characterized by significant limitations both in intellectual functioning and in adaptive behaviour (as cited in (Kleefstra and Hamel, 2005). This disability originates before the age of 18 years and the intellectual criterion for the diagnosis of mental retardation is represented by an intelligence quotient (IQ) of 70 or less (Lichten and Simon, 2007) (Fredericks and Williams, 1998). The importance of genes on the X-chromosome in the cause of mental retardation has been recognized for decades, due to two main considerations – males outnumber females in nearly all surveys of mental retardation by approximately a third (Lehrke, 1972); and numerous affected families have been observed in which mental retardation segregated in an X-linked inheritance pattern (Ropers and Hamel, 2005).

Traditionally, XLMR is divided into two categories (Ropers, 2006):

- Syndromic XLMR (S-XLMR), where mental retardation is associated with specific clinical, radiologic, or metabolic features, and
- Non-syndromic XLMR (NS-XLMR), where impairment of cognitive functions is observed without any other apparent features.

Both clinical and genetic evidence confirm that XLMR is a particularly heterogeneous group of syndromes, and allelic and locus heterogeneity is common in various clinical forms of XLMR (Ropers, 2006). For the more than 200 forms of XLMR described, 71 causative genes have been identified to date, summarised in Table 3.1 (De Brouwer et al., 2007; Ropers, 2006; Ropers and Hamel, 2005). Some of the identified genes have been implicated in both S-XLMR and NS-XLMR, which suggests that there may be no molecular basis for the segregation of the two categories (Ropers, 2006; Ropers and Hamel, 2005).

It can be seen in Table 3.1 that the identified genes are for the most part for syndromic forms of XLMR. The genetic heterogeneity of many forms of NS-XLMR greatly complicates the search for causative genes, as studies often need to be limited to studying isolated families (as the often clinically indistinguishable non-syndromic forms imply that one cannot pool linkage information from unrelated families). Wide linkage intervals and equally large numbers of candidate genes further complicate the elucidation of causative genes.

Linkage analysis has been performed for many forms of XLMR. In these cases the syndrome has successfully been mapped to a region on the X chromosome (Table 3.2), but the disease gene(s) not yet identified. Figure 3.1 illustrates where these linked regions are located on the X chromosome. It can be seen that these loci span most of the chromosome and many overlap, which implies that a single gene could potentially be involved in different forms of XLMR.

Currently numerous studies are underway to further elucidate the underlying genetic contributors of XLMR, and causative genes are continually identified.

Table 3.1: Genes implicated in XLMR. The genes are listed in the order of their location on the X chromosome. Compiled from (De Brouwer et al., 2007; Ropers, 2006; Ropers and Hamel, 2005). The Ensembl database listed 814 HGNC IDs for the X chromosome, of which 71 causative genes have been identified to date,

P ARM			Q ARM		
Gene	**Cytoband**	**XLMR form**	**Gene**	**Cytoband**	**XLMR form**
NLGN4X	Xp22.32-.31	NS-XLMR	*ARHGEF9*	Xq11.1	NS-XLMR
VCX	Xp22.31	NS-XLMR	*OPHN1*	Xq12	S-XLMR
MID1	Xp22.2	S-XLMR	*DLG3*	Xq13.1	NS-XLMR
PRPS2	Xp22.2	NS-XLMR	*NLGN3*	Xq13.1	NS-XLMR
OFD1	Xp22.2	S-XLMR	*SLC16A2*	Xq13.2	S-XLMR
NHS	Xp22.2-22.13	S-XLMR	*KIAA2022*	Xq13.2	NS-XLMR
CDKL5	Xp22.13	S-XLMR	*ZDHHC15*	Xq13.3	NS-XLMR
PDHA1	Xp22.12	S-XLMR	*ATRX*	Xq21.1	S-XLMR
RPS6KA3	Xp22.12	S-XLMR	*ATP7A*	Xq21.1	S-XLMR
SMS	Xp22.11	S-XLMR	*PGK1*	Xq21.1	S-XLMR
ARX	Xp22.11	S-XLMR/NS-XLMR	*TIMM8A*	Xq22.1	S-XLMR
IL1RAPL1	Xp21.2	NS-XLMR	*NXF5*	Xq22.1	NS-XLMR
GK	Xp21.2	S-XLMR	*PLP1*	Xq22.2	S-XLMR
DMD	Xp21.2-p21.1	S-XLMR	*PRPS1*	Xq22.3	S-XLMR
OTC	Xp11.4	S-XLMR	*ACSL4*	Xq23	NS-XLMR
TSPAN7	Xp11.4	NS-XLMR	*PAK3*	Xq23	NS-XLMR
BCOR	Xp11.4	S-XLMR	*DCX*	Xq23	S-XLMR
ATP6AP2	Xp11.4	S-XLMR	*AGTR2*	Xq23	NS-XLMR
MAOA	Xp11.3	S-XLMR	*LAMP2*	Xq24	S-XLMR
NDP	Xp11.3	S-XLMR	*GRIA3*	Xq25-26	NS-XLMR
EFHC2	Xp11.3	S-XLMR	*OCRL*	Xq25	S-XLMR
ZNF674	Xp11.3	NS-XLMR	*GPC3*	Xq26.2	S-XLMR
ZNF41	Xp11.3	NS-XLMR	*PHF6*	Xq26.2	S-XLMR
SYN1	Xp11.23	S-XLMR	*HPRT1*	Xq26.2	S-XLMR
ZNF81	Xp11.23	NS-XLMR	*ARHGEF6*	Xq26.3	NS-XLMR
FTSJ1	Xp11.23	NS-XLMR	*SOX3*	Xq27.1	S-XLMR
PQBP1	Xp11.23	S-XLMR/NS-XLMR	*FMR1*	Xq27.3	S-XLMR
SHROOM4	Xp11.22	S-XLMR	*AFF2*	Xq28	NS-XLMR
JARID1C	Xp11.22	S-XLMR/NS-XLMR	*IDS*	Xq28	S-XLMR
HSD17B10	Xp11.22	S-XLMR	*MTM1*	Xq28	NS-XLMR
PHF8	Xp11.22	S-XLMR	*SLC6A8*	Xq28	S-XLMR/NS-XLMR
FGD1	Xp11.22	S-XLMR/NS-XLMR	*ABCD1*	Xq28	S-XLMR
KLF8	Xp11.21	NS-XLMR	*L1CAM*	Xq28	S-XLMR
			MECP2	Xq28	S-XLMR/NS-XLMR
			FLNA	Xq28	S-XLMR
			GDI1	Xq28	NS-XLMR
			IKBKG	Xq28	S-XLMR
			DKC1	Xq28	S-XLMR

Table 3.2: A list of loci that have been linked to different forms of XLMR (listed in order of their location on the X chromosome). No causative gene(s) have as yet been identified for these linked regions. Syndromes that are listed as an entry on the OMIM database have the OMIM ID listed and where available the key reference mentioning the linkage analysis is listed. The numbers and colours are linked to the coloured lines used to indicate location in *Figure 3.1*

Nr	Locus	OMIM ID	PubMed ID
P ARM			
1	Xp22.3	%300406	(Dessay et al., 2002)
2	Xp22.1-p21.3		(Van Esch et al., 2005)
3	Xp22.13	#302350	(Zhu et al., 1990)
4	Xp22.13-p21.1	%300148	(Steinmuller et al., 1998)
5	Xp22	%304050	(Ballabio and Andria, 1992)
6	Xp22	%300421	(Wittwer et al., 1996)
7	Xp21.1-p11.22	%309610	(Watty et al., 1991)
8	Xp11.3-4	%300422	(Piluso et al., 2003)
9	Xp11.22	%309545	(Wilson et al., 1991)
Q ARM			
10	Xq11-q21	%300519	(Martin et al., 2000)
11	Xq12-q21	%300262	(Abidi et al., 1999)
12	Xq12-q21.31		(Shrimpton et al., 2000; Shrimpton et al., 1999)
13	Xq13-q22	%309605	(Miles and Carpenter, 1991)
14	Xq13.2-q21.2		(Stevenson et al., 1997)
15	Xq21.33-q23		(Chudley et al., 1999)
16	Xq22.3	%300581	(Jehee et al., 2005)
17	Xq23-q24		(Carpenter et al., 2000)
18	Xq24	*300360	(Vitale et al., 2001)
19	Xq25-q27		(Cilliers et al., 2007)
20	Xq26-q27	%300238	(Shashi et al., 2000)10677307
21	Xq26-q27	%307700	(Trump et al., 1998)
22	Xq26-q27.1	%304340	(Huang et al., 1991)
23	Xq27.3-q28	%302000	(Wijker et al., 1995)
24	Xq27-q28	%309800	(Graham et al., 1991)
25	Xq27-q28	%301590	(Graham et al., 1991)
26	Xq27-q28	%309620	(Dlouhy et al., 1987)
27	Xq28	%300261	(Armfield et al., 1999)
SPANNING			
28	Xp21.1-q22	%309585	(Wilson et al., 1991)
29	Xp21.2-q13		(Turner et al., 1994)
30	Xp11.3-q23	%300218	(Ahmad et al., 1999)
31	Xp11.1-q21.2		(Johnson et al., 1998)
32	Xp11.4-q24		(Oosterwijk et al., 1999)

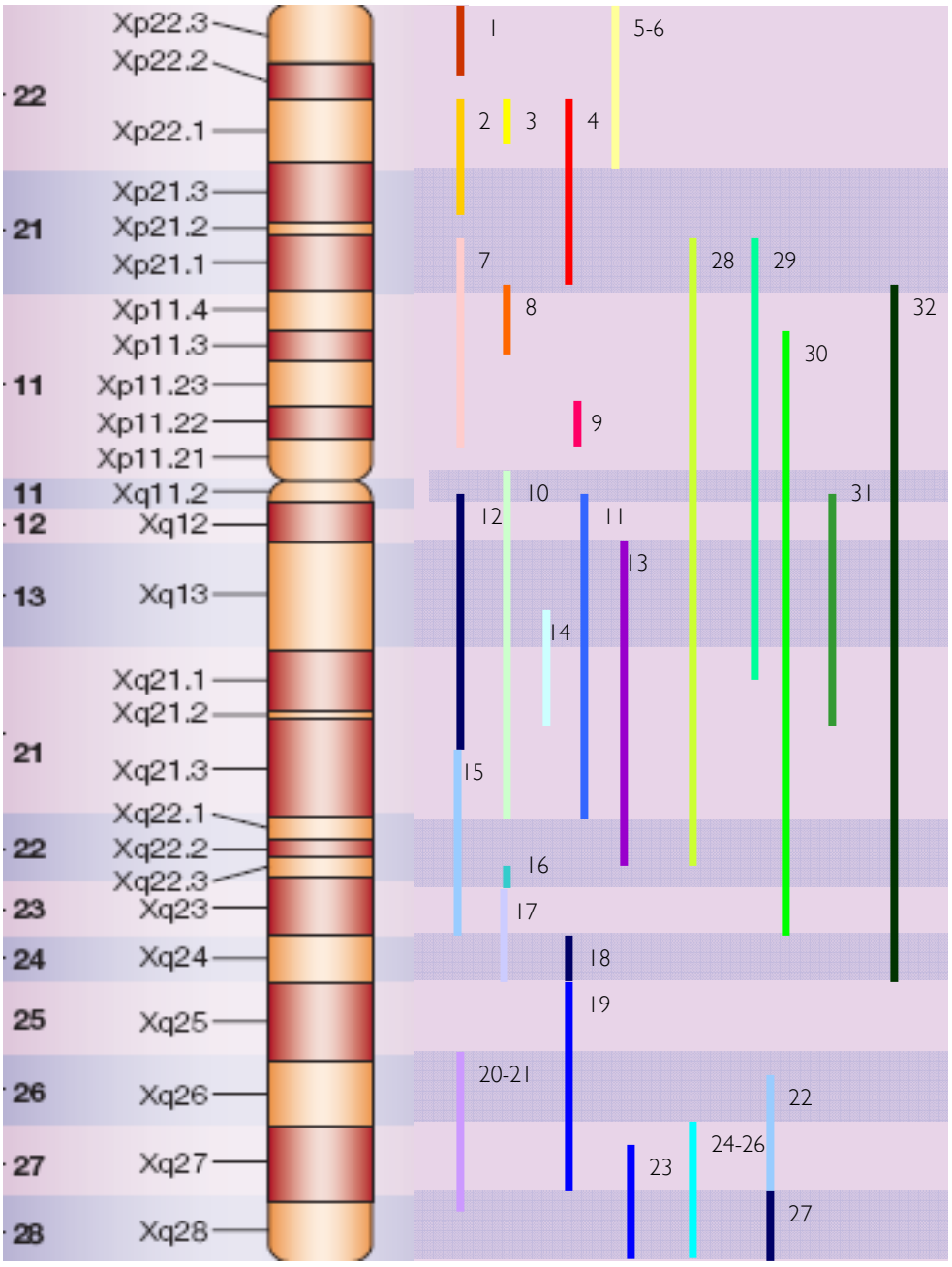

Figure 3.1: Regional localisation of different forms of XLMR. Line numbers and colours correlate to Table 3.2.

3.2 METHODS

3.2.1 Candidate gene list selection

A list of HGNC IDs for the genes on the X chromosome were obtained from the Ensembl database v48 (December 2007) and used as the initial candidate gene list that would be prioritized by the computational method. The Ensembl database listed 814 HGNC IDs for the X chromosome.

As described more specifically in Chapter 2, HGNC IDs are the identifiers preferably used here, as the HGNC is the global authority to assign unique and standardized gene symbols to human genes. This differs from the approach in Chapter 2, where the method described by Tiffin et al. (2005) was used to select a list of genes that could be submitted to binary filtering to prioritize the candidates. It was decided not to follow this approach here, because:

- Unlike the case with FASD, a genetic locus is defined (the X chromosome), and it is therefore not necessary to use the whole-genome as a starting point (as was the case with FASD)
- The selection process of Tiffin et al. (2005) focuses exclusively on anatomical sites related to the disease of interest. The binary filtering process allows for a more elaborate assessment, and anatomical sites related to the disease of interest were included as categories of criteria for evaluation, among other criteria points that were evaluated.

3.2.2 Selection of criteria for binary filtering

The computational method described in *Chapter Two* accesses various relevant database sources to establish whether a candidate gene, and its product, exhibit the biological characteristics expected to presume a link to a particular disease. Genes from the candidate gene list that were present in most criteria lists received the highest rank as putative candidates, with the premise that genes present in the largest number of independent criteria lists are least likely to be false positives. Four main categories of criteria were used for this evaluation – anatomical sites, phenotypes, biological processes related to the disease and homologous genes in model animal systems. Each criterion was used to populate a gene list, which was then used in the binary filtering process. To illustrate how such a gene list would be populated, consider anatomical sites as an example. Literature related to XLMR were collected and mined for anatomy ontology terms.

The anatomy terms found to be related to the disease were then used to extract lists of genes annotated with these terms from the Ensembl database v48 (December 2007). Other similar queries were formulated to extract gene lists for the other criteria lists, and in this way 40 criteria gene lists were populated (Table 3.3). A more elaborate description of the criteria categories and how the relevant gene lists were populated is given below:

Table 3.3: Summary of the criteria used to extract gene lists to compare to the candidate gene list (all X chromosome genes) to create a binary grid. A total of 40 criteria relevant to XLMR were used to populate the gene lists.

CRITERIA CATEGORIES			
Anatomical site	**Biological Process**	**Phenotype**	**Animal model homology**
Developmental	Development	Seizures	*Phenotype*
Liver	Transcription	Epilepsy	Behaviour/Neurological
Cns	Metabolism	Acidosis	Nervous system related
Respiratory	Phosphorylation	Microcephaly	Embryogenesis
Cerebellum	Brain development	Tremor	
Kidney			*Timing*
Hippocampus			Pre-Embryonic
Spinal cord			Embryonic
Cerebral cortex			Fetal
Testis			
Brain stem			*Anatomy*
Peripheral nerve			TS[1]8-9 Ectoderm
Cerebrum			TS10-13 Neural Ectoderm
Substantia nigra			TS14-26 CNS
Cardiovascular			
Adrenal gland			
Thyroid			
Ovary			
Amygdala			
Musculoskeletal			
Ganglion			
Hypothalamus			

[1]TS – Theiller stage: A term used to denote the stage of development of a mouse as described by Theiler in "The House Mouse: Atlas of Mouse Development" (Springer-Verlag, New York, 1989)

Anatomical site, phenotype and biological process

For all three these criteria categories, the scientific literature is used to obtain terms relevant to XLMR. The scientific literature is the data source that remains the most comprehensive source of disease-related information (Korbel et al., 2005). The scientific literature on XLMR was subjected to data-mining to obtain ontology terms related to

anatomical sites (eVOC terms), phenotypes (OMIM terms) and biological processes (GO terms) pertinent to XLMR.

Abstracts related to XLMR were obtained from the PubMed scientific literature database, using the following query: "(mental retardation, X-linked [MH])" Limits: only items with abstracts, English. The online literature mining tools DDE and DTFAM were used to extract the relevant ontology terms from the body of literature, whereafter gene lists were extracted from the Ensembl database, as described before, for binary evaluation purposes.

Animal model homology

Animal models offer major contributions to the understanding of human disease. Although many different animal models for XLMR have been developed, the mouse model is the model most often used (Watase and Zoghbi, 2003). Three areas of homology were chosen as criteria in this category:

- Genes associated with phenotypes affected by the disease
- Genes expressed at different developmental stages
- Genes expressed in the developing brain

MGD was accessed to select genes associated with the above-mentioned categories (Blake et al., 2006; Eppig et al., 2005). Human homologues were identified and used to populate these criteria lists.

3.2.3 Binary filtering and prioritization of genes on the X chromosome

For each of the genes on the X chromosome a binary score was calculated simply by summing all binary scores for each of the criteria lists used. This was achieved by comparing the criteria-specific gene lists generated to the candidate gene list to create a binary matrix.

The binary evaluation was performed as follows: A gene in the candidate gene list was assigned a 1 when that gene was also present in the criteria gene list that it was evaluated against. If the gene was absent it was assigned a 0. All genes were then ranked based on this score, with those having higher scores being higher in the rank list. Genes in the candidate list that were present in most criteria lists (obtaining the most 1-scores in the binary matrix) received the highest rank as candidates.

3.2.4 Evaluation of biological significance of prioritized genes for XLMR

It was expected that the described computational method will not be 100% accurate (i.e. detect all known XLMR genes), but that the top-ranked part of the candidate gene list will be significantly enriched for known and presently unknown XLMR causative genes. It is therefore possible that some of the top-ranked genes that are not currently known XLMR genes may be good candidates for XLMR syndromes for which the genetic cause has not been identified. This anticipated link between XLMR and the top-ranked genes was evaluated by assessing protein-protein interactions and by functional enrichment analysis.

It is possible that top-ranked genes that are not known XLMR genes are false-positives. To evaluate the likelihood of the non-XLMR top-ranked genes being XLMR candidates, functional annotation (GO terms) of the non-XLMR top-ranked genes was compared to that of known XLMR genes.

- **Protein-protein interactions**

The biological significance of the prioritized genes was evaluated by using STRING v7.1 to assess protein-protein interaction among the top-ranked genes (Von Mering et al., 2007). The STRING database provides a comprehensive source of protein-protein association evidence under a common framework. The top-ranked genes and randomly selected low-ranked genes were used as input and a medium confidence score for evidence was selected (50%).

- **Functional enrichment analysis**

The online tool DAVID 2007 (Dennis et al., 2003) was used to evaluate if significant functional enrichment is present among the top-ranked gene list. DAVID is an online tool that integrates genomic functional annotations to reveal biologically relevant enrichment in a gene list. DAVID promotes functional discovery through exploration of biochemical pathway maps, functional classification using GO terms and conserved protein domain architecture. Data from various sources are integrated into DAVID, including GenBank, UniGene, RefSeq, Locuslink, KEGG, OMIM and GO. The top-ranked genes were submitted as a list to DAVID 2007 (January 2007), which was then compared to a background gene list to assess functional enrichment within the list. The background list is user-defined – for this analysis two runs with different background lists were performed.

Firstly, all genes on the X chromosome were used as background, and secondly the *Homo sapiens* default background list from the DAVID website was used as background.

For illustrative purposes, only genes ranked within the top 10 criteria matched categories will be used for evaluation of biological significance. This is an arbitrary cut-off used to select an appropriately sized group of prioritized genes to investigate, which could typically also be suitable for a candidate gene association study. A negative control set was generated to assess the validity of the ranking method, by randomly selecting a set of genes of the subset matching to no criteria.

3.2.5 Identifying candidate genes for XLMR with unknown genetic aetiology

An assessment was made of whether genes from the regions shown to be linked to XLMR (Table 3.2 and Figure 3.1) were among the top-ranked genes. If genes from these linked regions are prioritized by the binary filtering method, they become putative candidates for the forms of XLMR for which the genetic cause has not yet been uncovered.

3.3 RESULTS

The binary filtering process described in Chapter 2 was evaluated to assess efficacy and accuracy, by applying it to prioritize candidate genes for XLMR. Furthermore the method was applied to uncover unknown candidate genes and emerging pathways for XLMR.

3.3.1 Binary prioritization of XLMR genes

Table 3.4 summarises the prioritization of XLMR genes by the binary filtering process. The gene with the most criteria matched (27 matches) most was *MECP2*, a known XLMR causative gene. Mutations in *MECP2* are known to cause Rett syndrome (a syndromic form of XLMR, OMIM: #312750), as well as non-syndromic male fatal neonatal encephalopathy, progressive spasticity and non-syndromic Angelman and Prader–Willi-like phenotypes (Amir et al., 1999; Meloni et al., 2000; Renieri et al., 2003).

Table 3.4: Prioritization of genes on the X chromosome by the binary filtering process.The enrichment for known XLMR genes is shown. Approximately half of all known XLMR genes were in the top 16 matched categories (indicated by dashed line). The double-line indicates the top 10 matched categories – genes in these categories were used for further analysis.

Nr of criteria matched	Genes in matched category	XLMR genes in matched category	Other genes in matched category	% detection of XLMR genes
27	1	1	0	100
24	1	1	0	100
23	1	1	0	100
22	1	0	1	0
21	3	0	3	0
20	2	2	0	100
19	4	3	1	75
18	6	1	5	17
17	15	3	11	21
16	18	2	15	12
15	10	2	8	20
14	24	3	21	13
13	21	4	17	19
12	35	3	32	9
11	31	5	26	16
10	28	6	22	21
9	33	4	29	12
8	29	3	26	10
7	36	3	33	8
6	23	2	21	9
5	35	4	31	11
4	45	2	43	4
3	47	5	42	11
2	80	7	73	9
1	96	4	92	4
0	189	0	189	0

Approximately half (52%) of all known XLMR genes were in the top 16 matched categories (indicated by the dashed line in Table 3.4). The top-half of the prioritized list (represented by a cut-off of four criteria matched, containing 402 of the 814 X chromosome genes) contains 55 of the known 71 XLMR genes – denoting an enrichment of 78%. The enrichment of each matched category becomes more moderated as one progress down the ranked list, indicating an efficient ranking of known disease genes over less likely candidates.

Genes that matched one or none of the criteria are considered to be less likely to be associated with XLMR. Based on this premise, 285 genes (35%) from the candidate gene list were ranked as weak candidates. There were four XLMR genes present in this category – all four genes matching only one criterion. These four genes are *NXF5*, *PRPS2*,

ZDHHC15 and *NHS*. With the exception of *NHS* all these genes have been found to be involved in different forms of NS-XLMR.

For the purpose of evaluating putative biological enrichment among the prioritized candidates, genes matching to the top 10 criteria matched categories were used (cut-off indicated by double line in Table 3.4). Based on this premise a list of 52 genes was selected, containing 14 known XLMR genes.

3.3.2 Evaluation of biological enrichment among prioritized genes for XLMR

It is expected that other top-ranked genes that are not known candidates, may be false-positives, or as yet unknown candidate genes for forms of XLMR with unknown genetic aetiology. To evaluate the likelihood of the non-XLMR top-ranked genes being XLMR candidates, functional annotation (GO terms) of the non-XLMR top-ranked genes was compared to that of known XLMR genes. Only GO terms that were significantly enriched within the two lists (as analysed using DAVID) were used for the analysis. It was observed that these two gene lists shared several GO terms, implying a functional similarity of the two lists (XLMR genes and top-ranked non-XLMR).

By investigating the functional enrichment of the prioritized genes, one can assess which of these genes are the best candidates for XLMR. This is achieved by: (a) Assessing protein-protein interaction among the top genes, and (b) observing significant enrichment of biologically relevant annotation (pathways, GO terms and shared protein domain terms) among the top genes. The following functional links were observed among the top-ranked genes for XLMR:

- **Protein-protein interactions**

Both the top-ranked and low-ranked genes were submitted to the STRING database (von Mering et al., 2005) to assess known protein-protein interactions. For the genes that were found to be linked through protein-protein interaction, the source of evidence for the interactions and confidence scores are summarized in Table 3.5. Figure 3.2 shows the STRING network of interactions. The network view summarizes the associations for the group of gene products. The network edges represent the predicted functional associations and each colour represents a different line of evidence. No known interactions were observed among the low-ranked genes.

Table 3.5: Known protein-protein interaction for the prioritized genes obtained using STRING. The available evidence for the most significant interactions as well as the confidence score assigned for the interactions are shown.

Gene 1	Gene 2	Confidence scores		Combined confidence score
		Experimental evidence	Textmining co-occurrence	
ATRX	MECP2	0	0.920	0.920
ARAF	SYN1	0	0.873	0.873
SYN1	TIMP1	0	0.848	0.848
ARAF	TIMP1	0	0.799	0.799
PCTK1	UBE1	0	0.742	0.742
GPM6B	MECP2	0	0.724	0.724
BGN	TIMP1	0	0.671	0.671
PCTK1	USP11	0	0.654	0.654
A15	SYN1	0	0.618	0.618
UBE1	SYN1	0	0.61	0.610
ATRX	A15	0	0.595	0.595
MECP2	A15	0	0.561	0.561
SLC6A8	A15	0	0.533	0.533
UBE1	USP11	0	0.531	0.531
HDAC6	USP11	0.526	0	0.526
UBE1	MECP2	0	0.506	0.506
SLC6A8	ATRX	0	0.497	0.497
UBE1	ARAF	0	0.497	0.497
NONO	HTATSF1	0	0.494	0.494
MIC5	PLP1	0	0.488	0.488
MIC5	MECP2	0	0.477	0.477
A15	PQBP1	0	0.467	0.467
MAGED1	NGFRAP1	0	0.463	0.463
SYN1	USP11	0	0.429	0.429
ARAF	USP11	0	0.413	0.413
DMD	EIF2S3	0	0.411	0.411
MIC5	HPRT1	0	0.407	0.407
HPRT1	UBE1	0	0.407	0.407
HPRT1	IDS	0	0.402	0.402

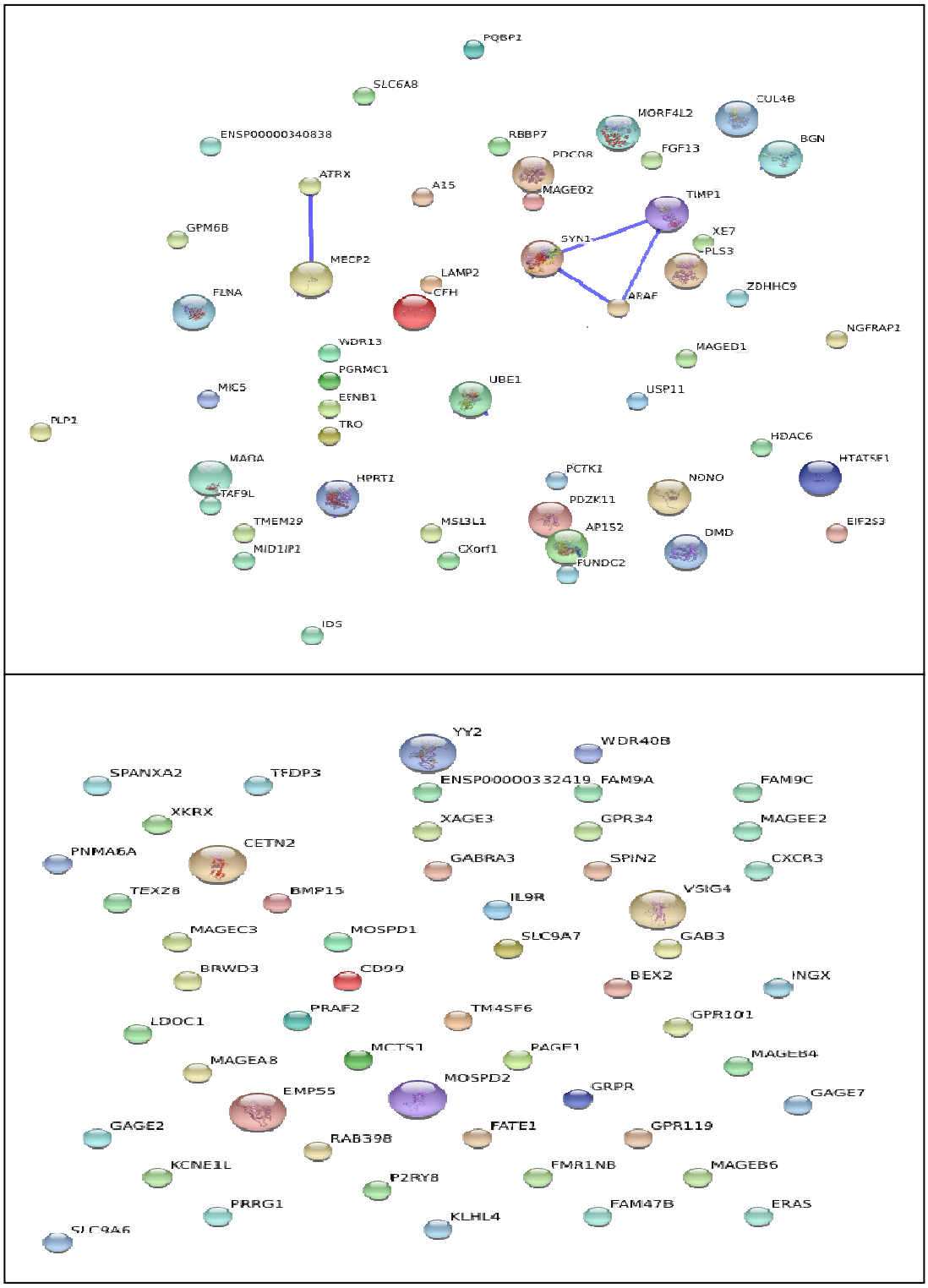

Figure 3.2: The STRING network of known protein-protein interactions among the (a) 50 top-ranked candidate genes for XLMR and (b) 50 randomly selected bottom-ranked genes. The network edges represent the predicted functional associations. The network edges represent the predicted functional associations, with the thickness of the line representing the confidence of the predicted interaction.

- **Functional enrichment analysis**

DAVID 2007 was used to assess functional enrichment within the top-ranked candidate gene list. Firstly, significant enrichment of GO terms and conserved protein domain architecture was observed in the top-ranked list. The GO terms found to be significantly enriched for the top-ranked gene list are shown in Table 3.6. Concerning conserved protein domain architecture; it was observed that the myelin proteolipid protein motif featured prominently among the top-ranked genes. Furthermore actin-binding proteins were another significant feature motif among top-ranked genes.

Table 3.6 Biological process and molecular function GO terms significantly over-represented among the top-ranked genes. The gene count (N) indicates how many genes from the list were annotated with the particular GO term. Note that varying *P*-values were obtained depending on the background list used

GO Term	N	*P*-Value[1]	*P*-Value[2]	Top-ranked genes with this feature
Development	16	7.24×10^{-6}	3.9×10^{-4}	*FLNA, MORF4L2, RBBP7, PLP1, EFNB1, GPM6B, FHL1, TIMP1, HDAC6, DMD, MSL3L1, NGFRAP1, TRO, FGF13, ATRX, L1CAM*
Nervous system development	6	0.005	0.013	*FLNA, EFNB1, PLP1, FGF13, L1CAM, GPM6B*
System development	6	0.005	0.015	*FLNA, EFNB1, PLP1, FGF13, L1CAM, GPM6B*
Regulation of neurotransmitter levels	3	0.005	0.014	*SYN1, SLC6A8, MAOA*
Cell differentiation	6	0.006	0.003	*FHL1, TIMP1, EFNB1, PLP1, L1CAM, GPM6B*
Cell organization and biogenesis	10	0.007	0.061	*FHL1, TIMP1, AP1S2, HDAC6, DMD, FLNA, MSL3L1, MORF4L2, MID1IP1, ATRX*
Cell-cell signalling	6	0.008	0.002	*SYN1, SLC6A8, EFNB1, PLP1, MAOA, FGF13*
Organelle organization and biogenesis	7	0.013	0.076	*HDAC6, DMD, FLNA, MSL3L1, MORF4L2, MID1IP1, ATRX,*
DNA metabolism	6	0.020	0.061	*NONO, HDAC6, MSL3L1, MORF4L2, UBE1, ATRX*
Transmission of nerve impulse	4	0.021	0.005	*SYN1, SLC6A8, PLP1, MAOA*
Synaptic transmission	3	~	0.044	*SYN1, SLC6A8, MAOA*
Chromosome organization and biogenesis	4	0.030	~	*HDAC6, MSL3L1, MORF4L2, ATRX*
Chromatin modification	3	0.035	~	*HDAC6, MSL3L1, MORF4L2*
Chromosome organization and biogenesis	4	0.035	~	*HDAC6, MSL3L1, MORF4L2, ATRX*
Cell maturation	2	0.079	~	*TIMP1, PLP1*

[1]P-value obtained using the Homo sapiens gene list as a background list to the top-ranked candidate genes; [2]P-value obtained using all genes on the X chromosome as a background list to the top-ranked candidate genes

Furthermore, pathway information was analysed, highlighting the axon guidance pathway as being represented among the top-ranked genes, although the enrichment was only marginally significant. Only two genes from the top-ranked gene list were members of this pathway (*EFNB1* and *L1CAM*).

3.3.3 Identifying candidate genes for XLMR with unknown genetic aetiology

It is likely that the top-ranked gene list will be enriched with as yet unknown candidate genes for XLMR. Therefore the top 50 genes were compared to genes present in the loci that have been linked to different forms of XLMR with no known cause. There was significant overlap between the top-ranked genes for XLMR and genes from the XLMR linked regions. A table matching the top-ranked genes to the linked regions mentioned in Table 3.2 is shown in *Addendum B*.

There were five genes that matched to nine of the linked regions – *ATRX, EFNB1, NONO, PDZD11* and *TAF9B*. Although the XLMR forms that are linked to these regions all have various phenotypic characteristics associated with them, the majority of the syndromes have skeletal anomalies, such as microcephaly, in common. Table 3.7 summarises the genomic and expression information for these putative candidates, whereas their location is indicated in Figure 3.3.

Table 3.7: XLMR candidate genes matching to most XLMR linked regions.

	HGNC ID	Location	Description	Brain expression?	Fetal expression?
1	ATRX	Xq13.1-21.1	ATP-dependent helicase	Y	Y
2	EFNB1	Xq12	Eph-related receptor tyrosine kinase ligand 2	Y	Y
3	NONO	Xq13.1	Nuclear RNA-binding protein	Y	Y
4	PDZD11	Xq13.1	PDZ domain-containing protein 11	Y	Y
5	TAF9B	Xq13.1-21.1	Transcription initiation factor TFIID subunit 9B	Y	Y

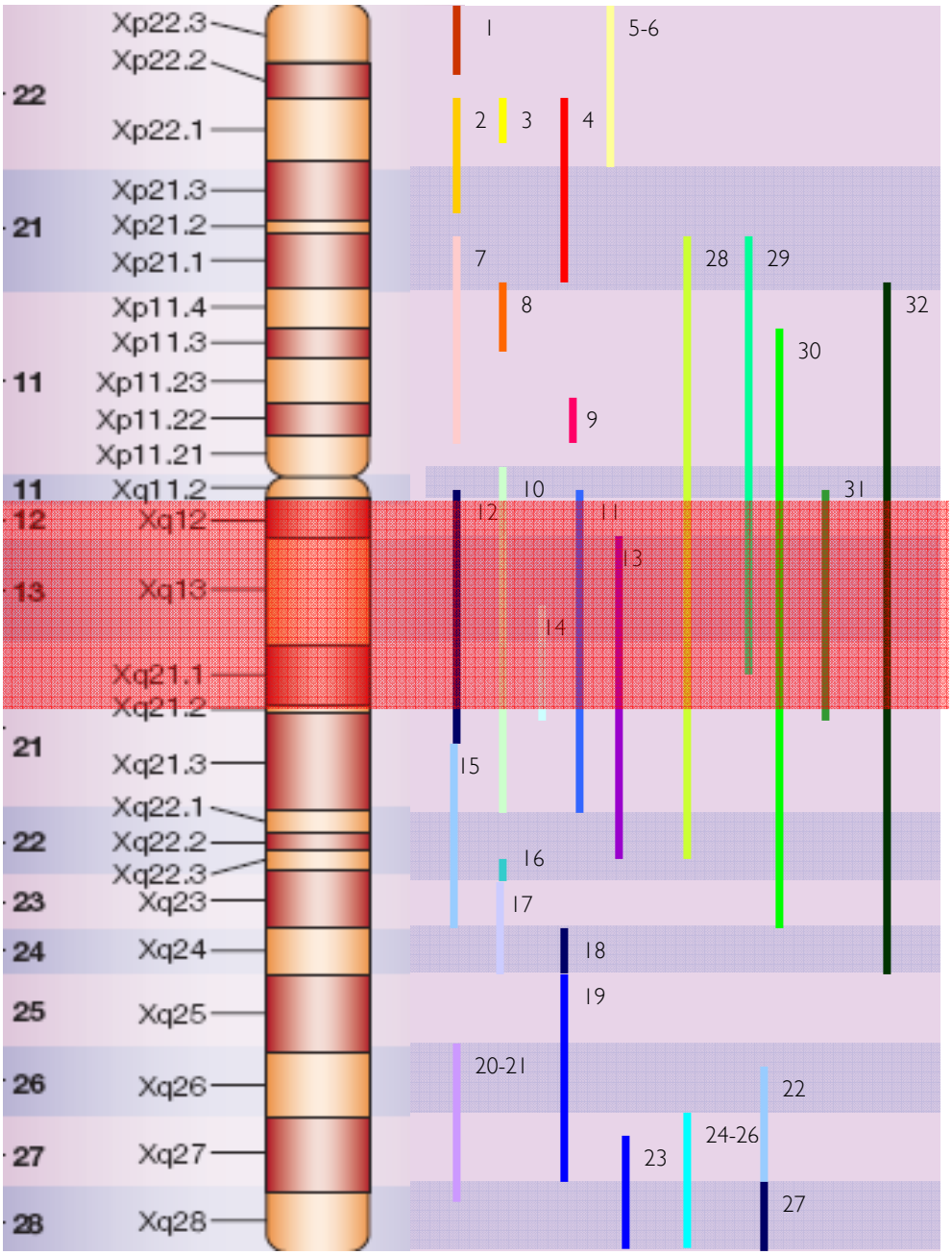

Figure 3.3: Heat map indicating the regional localisation of five top-ranked genes and their location in comparison to the different forms of XLMR with unknown etiology. The chromosomal region highlight by the red block is where the five candidates are concentrated. Line numbers and colours correlate to Table 3.2.

3.4 DISCUSSION

The aim of this study was to assess the effectiveness of the described computational candidate gene prioritisation method in selecting appropriate candidates. The binary filtering approach was evaluated on XLMR, a set of heterogenous disorders, of which some of the underlying genetics is known. The computational method was therefore applied to evaluate the effectiveness of the approach in prioritizing known XLMR genes, and at the same time uncover unknown candidate genes and emerging pathways for XLMR.

3.4.1 Binary prioritization of XLMR genes

Applying the binary filtering approach to select XLMR genes resulted in a prioritized gene list with a noted enrichment of known XLMR genes among the top-ranked genes (Table 3.4).

The gene with the most criteria matched (27 matches) most was *MECP2*, a known XLMR causative gene. Mutations in *MECP2* are known to cause Rett syndrome. *MECP2* is a widely expressed transcriptional repressor. *MECP2* has two conserved functional domains, the methyl- CpG binding domain and the transcription repression domain. *MECP2* displays extreme allelic heterogeneity, with more than 100 different mutations in the *MECP2* gene being described in patients with Rett Syndrome (Renieri et al., 2003). Further to this, *MECP2* mutations have also been shown to produce non-syndromic male fatal neonatal encephalopathy, progressive spasticity and non-syndromic Angelman and Prader–Willi-like phenotypes (Amir et al., 1999; Meloni et al., 2000; Renieri et al., 2003).

Rett syndrome is a prime example of the locus heterogeneity associated with XLMR *MECP2* mutations account for only approximately 70–80% of cases, whereas locus heterogeneity is hypothesized to explain the occurrence of the syndrome among *MECP2* negative cases (Renieri et al., 2003). It is therefore likely that other genes that have been prioritized by the binary ranking method may also be genetic candidates for Rett syndrome (amongst others), whereas *MECP2* may have an effect on other forms of XLMR with yet unknown genetic aetiology.

This approach is primarily based on gene annotation, and the key consideration is therefore the selection of appropriate criteria for binary evaluation. The main limiting factor of the described computational approach is therefore the data contained in the public domain that

is used for the binary evaluation. It is feasible that the method has a biased probability of prioritizing better annotated genes over those that are not as well annotated, regardless of whether these genes are relevant to the disease of interest. The selection of appropriate criteria and data sources are essential requirements, and critical to valid candidate gene selection.

3.4.2 Evaluation of biological enrichment among prioritized genes for XLMR

Evaluation using available online tools such as DAVID and STRING highlighted protein-protein interaction, functional enrichment and probable biological significance among the top-ranked genes. An analysis of the functional enrichment among the top-ranked genes, give a reliable indication of the appropriateness of these genes as XLMR candidates. STRING is a useful tool to highlight genes within the top-ranked gene list that interact and that may have a cooperative effect on disease outcome. Data obtained with STRING indicate that there is a solid interaction grid among the top-ranked genes (Table 3.5 and Figure 3.2), indicating that there is underlying molecular pathways that have genetic contributors that influence mental functioning.

DAVID elucidates functional enrichment and biological significance within the top-ranked gene list, and highlighted the axon guidance pathway as a central starting place for investigation. Axon guidance is of importance in CNS development and cognition, and malfunction of gene products from this pathway could play a possible key role in XLMR pathogenesis (Thiery, 2003; Walsh and Doherty, 1997). Furthermore, DAVID highlighted various biological processes that are important for XLMR, and indicated which of the prioritized genes have been annotated with these functions (Table 3.6).

3.4.3 Identifying candidate genes for XLMR with unknown genetic aetiology

The top 50 genes were compared to genes present in the loci that have been linked to different forms of XLMR with no known cause. There was significant overlap between the top-ranked genes for XLMR and genes from the XLMR linked regions. There were five genes that matched to nine of the linked regions – *ATRX, EFNB1, NONO, PDZD11* and *TAF9B*. These five genes are highlighted as critical candidates to include in an experimental investigation of XLMR association, as they also have a significant number of the functional characteristics highlighted in the interaction- and enrichment analysis performed using DAVID and STRING.

A focused investigations of these five genes, showed these candidates to be particularly appropriate candidates for these regions, and for XLMR overall. Apart from the features mentioned in Table 3.7, such as expression in the brain and during embryonic development, there are some tentative links between these five genes and brain function. *ATRX* is a known XLMR causative gene, linked to X-linked alpha-thalassemia/mental retardation syndrome (OMIM: # 301040). The protein encoded by this gene belongs to the SWI/SNF family of chromatin remodelling proteins, and mutations have proven to cause changes in the pattern of DNA methylation, which may provide a link between chromatin remodelling, DNA methylation, and gene expression in developmental processes. It is known that ATRX is crucial for normal development and organization of the cortex, and emphasize the relevance of this gene as a candidate for other forms of XLMR as well (Berube et al., 2002).

EFNB1 is a critical candidate, as its gene product is involved in CNS development, angiogenesis, and neural synapses formation and maturation, as well as axon guidance, which was highlighted as an central pathway in XLMR by DAVID analysis (Han et al., 2002). Moreover, the EFNB1 gene have been shown to be involved in normal morphogenesis of skeletal elements, and mutations within the gene have been correlated to craniofrontonasal syndrome (Compagni et al., 2003; Wieland et al., 2004). This observation can possibly be extrapolated to the association of craniofacial anomalies associated with the XLMR syndromes that are linked to the regions overlapping with the location of *EFNB1.*

PDZD11 is a novel protein that regulates brain copper homeostasis (Stephenson et al., 2005). Aberrant copper homeostasis is implicated in neurodegenerative disorders such as Alzheimer disease and this suggests that impairment of copper efflux could negatively influence neuronal function and in this way contribute to rapid neuronal degeneration (Madsen and Gitlin, 2007; Schlief and Gitlin, 2006).

The link between *NONO* and *TAF9B* and XLMR is not as apparent as with the other candidates. Both these genes are transcriptional regulators, and could potential have a regulatory effect on another gene product that exerts an influence on XLMR development (Chen and Manley, 2003; Ishitani et al., 2003).

This preliminary investigation showing a tentative hypothesis as to the candidacy of these genes for XLMR leads to the suggestion that other top-ranked candidate genes may also be appropriate candidate genes for the particular linked region that they are associated with.

3.5 CONCLUSION

It would be difficult to describe the multitude of different biological and biochemical mechanisms simultaneously operating in each cell. This approach of computational candidate gene selection and prioritization provides an appropriate approach to select and focus on particular aspects related to a particular disease, and hereby identify a feasible candidate gene list for empirical investigation. The true test for the appropriateness of this technique would be to establish a statistically significant association between variations within these genes and XLMR, which can only be done experimentally. However it has been shown that the genes identified using this technique are biologically relevant to the disease and therefore appropriate for use in a candidate gene association study.

The validation process applied here illustrated that the described computational approach can be employed to select and refine an enriched candidate gene list by filtering the list based on criteria pertinent to the disease of interest. Using this approach resulted in a prioritized candidate gene list that was enriched with known XLMR genes and contained other top-ranking genes that are good candidates for experimental validation. The link between X chromosome genes and XLMR is not directly apparent in all cases, which highlights the main advantage of using a computational approach for candidate gene identification, which is to identify non-intuitive candidates. It should be considered that the approach described can select and refine a “most likely” candidate gene list according to known characteristics of the disease.

A final consideration is that, although far more is known about the role of X-linked genes in mental retardation, it is expected that autosomal genes and even structural variation also contribute to mental functioning. Furthermore, the noted allelic- and locus heterogeneity of XLMR, suggest that it is likely that risk factors exist for mental retardation that predispose to, but do not cause the phenotype (Ropers and Hamel, 2005). A future application of the computational process described here to identify and prioritize candidate genes for XLMR, can be expanded to evaluation of the whole-genome for mental retardation candidate genes.

Chapter 4

Analysis of genetic variation in *FGFR1* – a candidate gene for FASD

4.1 INTRODUCTION

A computational approach based on ranking of genes using binary filtering was taken to select a group of candidate genes to be tested for susceptibility to FASD, in the absence of genetic linkage evidence. This approach identified *FGFR1* as the top-ranked candidate gene, with other genes from the TGF-β, MAPK and Hedgehog signalling pathways also classified as likely being linked to disease susceptibility. The link between variation within these candidate genes and risk of FASD development can be evaluated by genotyping specific genetic variations in a group of affected and unaffected individuals, and subsequently performing statistical analyses to determine association. However, the power to discover a relationship between genetic variation and phenotype is limited by the sensitivity and accuracy of the methods available to detect said variation. An assortment of variation exists within the human genome, ranging from large, microscopically visible chromosome anomalies to single nucleotide changes (SNPs), as well as extra-nucleotide changes such as epigenetic phenomena. All categories of genetic variation could potentially influence human phenotype, and may therefore be investigated for disease risk involvement.

4.1.1 Human Genetic Variation

- **Copy number variation**

The human genome have been shown to contain various copy number variations (CNVs) of DNA segments, including deletions, insertions and duplications of assorted genomic segments. A CNV can be defined as a DNA segment that is 1 kb or larger and present at variable copy number in comparison with a reference genome. A CNV can be simple in structure, such as tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome. At present, the two main comparative techniques that are most appropriate to assess human CNV are genome-wide SNP arrays (such as the Affymetrix® Human SNP Array 6.0) and comparative genomic hybridization (CGH) arrays (of which the NimbleGen human CGH microarray platform is an example) (Redon et al., 2006). Although increasing evidence supporting the important contribution of CNV to the genetics of complex disease, knowledge of the fine-scale architecture of the majority of CNVs in the human genome is still incomplete (Perry et al., 2008), and few methods have been developed for the analysis of such variation in the context of genetic association studies (Ionita-Laza et al., 2008), complicating the routine use of these variations in disease-gene association studies.

- **Epigenetic variation**

The term Epigenetics refer to heritable variations that influence gene expression, with no change to the DNA sequence (Jirtle and Skinner, 2007). Known epigenetic mechanisms include DNA methylation, regulation by non-coding RNAs, chromatin remodeling, and silencing of repetitive, transposable elements. Epigenetic variation is a known key regulator of mammalian development, and has been linked to several human diseases, including Rett syndrome (Shahbazian and Zoghbi, 2002), Angelman- and Prader-Wili syndrome (Knoll et al., 1989), and several human cancers (Esteller, 2008; Moss and Wallrath, 2007; Risch and Plass, 2008). The importance of understanding epigenetic phenomena and its role in disease was highlighted by the establishment of the Human Epigenome Project, with the aim of identifying, cataloguing and interpreting genome-wide DNA methylation patterns of all human genes in all major tissues (Eckhardt et al., 2004).

Many novel techniques to study epigenetic variation have been developed including chromatin immunoprecipitation (ChIP) to assess chromatin binding (DeAngelis et al., 2008), synthetic peptide arrays (SPOT blot analysis) to assess histone modification (Nady et al., in press), and DNA Methylation Analysis (DeAngelis et al., 2008; Toyota et al., 1999). DNA methylation analysis is the most widely applied technique, in light of the importance of DNA methylation in development and disease, and due to the fact that methylation is the only flexible genomic parameter that can change genome function under exogenous influence. DNA methylation constitutes the main link between genetics, disease and the environment that is widely thought to play a critical role in the aetiology of many human diseases (Jirtle and Skinner, 2007). Variation in DNA methylation patterns is usually assessed by firstly treating DNA by bisulfite modification. Bisulfite modification of DNA converts cytosine into uracil, whereas 5-methylcytosine remains unchanged, essentially creating a C/T polymorphism. Consequently, the methylation status of a particular cytosine can be quantitatively assessed through direct sequencing (Shiraishi and Hayatsu, 2004), pyrosequencing (Tost and Gut, 2007) or massively-array genotyping (Bibikova et al., 2006).

The role of environmental exposure in epigenetic variation and disease is a particularly important consideration for the study of the susceptibilty of FASD development. Alcohol has been shown to lower levels of DNA methyltransferases, which are key enzymes that mediate the predominant epigenetic phenomenon of DNA methylation (Bielawski et al., 2002; Garro et al., 1991). Further to this, observations that paternal alcohol consumption may negatively influence fetal neurobehavioural development and growth (Jamerson et al.,

2004), also hint at a possible epigenetic mechanism in FASD development, and will need further investigation.

- **Single nucleotide polymorphisms**

As far as genetic variation goes, SNPs represent one of the most powerful markers utilised in the search for disease susceptibility genes. It is estimated that more than 10 million SNPs are present in the human population, about 1 every 300 nucleotides in any one individual's genome (Kruglyak and Nickerson, 2001). SNPs differ from other variations (like microsatellites and short tandem repeats) in that they are relatively easy to detect (Kim and Misra, 2007; Kwok, 2001; Nowotny et al., 2001) and are amenable to high throughput, low cost, automated typing (Dearlove, 2002; Ji et al., 2004; Nordfors et al., 2002; Sun and Guo, 2006; Tindall et al., 2007), making them the markers of choice in many association studies, and will also be the focus of the current investigation towards the putative association between variation within specific candidate genes and risk of FASD development.

Any candidate region or gene under study is likely to contain multiple SNPs which can be genotyped in an association study. However, typing all known SNPs in a candidate gene could be a laborious and costly exercise. A cost-saving approach would therefore be to select a subset of SNPs, selecting variants based on their putative functional effect, LD and haplotype blocks and the variant's distinctiveness and prevalence in the study population.

4.1.2 SNP selection based on function

SNPs that fall within the protein coding region of genes can affect the protein both qualitatively and quantitatively. Both synonymous (a variant that does not alter the polypeptide sequence) and non-synonymous (a variant that does alter the polypeptide sequence) SNPs can affect function in various ways.

The most apparent way in which a SNP can influence the function of a gene product is by altering an amino acid (missense mutation) or introducing a premature stop codon (nonsense mutation). SNPs can also impact on function by influencing exonic splicing enhancer (ESE) sequences and exon-intron boundaries. Non-coding SNPs can have a severe effect on the structure of the encoded protein, by inactivating an ESE, resulting in exon skipping and an altered protein product. ESEs also appear to be especially important in exons that normally undergo alternative splicing. Many point mutations linked to genetic

disease cause irregular splicing by disrupting splicing directly or by obstructing splicing regulatory element binding (Cartegni et al., 2002). SNPs within the exon-intron boundaries of a gene can further promote missplicing events.

Whilst the number of protein-coding genes in the human genome is much lower than first anticipated, control of gene expression is complex. These observations have led to the hypothesis that regulatory SNPs (rSNPs) are possibly more important in human disease phenotype due to the effect that they could exert on gene expression control (Buckland, 2004; Knight, 2005; Morley et al., 2004).

The core promoter region of a gene contains the elements essential to transcription initiation, and is therefore an obvious region to include when searching for putative functionally relevant variants. Furthermore, the 5' and 3' UTRs of a gene have been shown to be involved in many post-transcriptional regulatory pathways that control mRNA localization, stability and translation efficiency, implying that SNPs within this region may also have a functional effect (Buckland, 2006).

Computationally, the effect of an rSNP within the promoter region can be predicted by assessing whether the variant falls within a known or possible regulatory motif, such as a TFBS, and whether the position is conserved in model organisms.

4.1.3 SNP selection based on LD and haplotype blocks

The International HapMap Project was launched in October 2002, with the main objective to determine the common patterns of DNA sequence variation in the human genome (International HapMap Consortium, 2005; 2007). This was achieved by describing SNPs, their frequencies, and correlations between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. One of the major benefits of the project is to aid in the design of association studies, under the premise that SNPs within a haploptype block (i.e. the specific set of alleles observed on a single chromosome or part thereof that are in strong LD) are correlated. The co-occurrence of SNPs in haplotype blocks result in associations between these variants in populations (known as LD). This implies that genotyping of a selection of SNP alleles (now known as tagSNPs) in a particular region would provide sufficient information to predict most of the information about the genetic variation in that region. Recently, the outcomes of the second phase of HapMap was published, with more than 3.1 million human single nucleotide polymorphisms (SNPs)

genotyped in 270 individuals from four geographically diverse populations. It included 25–35% of common SNP variation in the populations surveyed (International HapMap Consortium , 2007).

One important consideration, however, is that the range and distribution of LD in different populations is largely unknown and haplotype blocks are population specific (International HapMap Consortium, 2005). The HapMap project contains information on a very limited number of populations, and can therefore not be extrapolated to all populations globally. Furthermore, it is thought that the effect of genotype on disease could vary between disorders and populations due to genetic and environmental heterogeneity (Weiss and Terwilliger, 2000). For these reasons haplotype block distribution was not the primary consideration for the selection of SNPs for future association studies for FASD, as the study population is an admixed group with parental contributions from indigenous South African, European and Asian populations, where the first is not represented in major haplotype databases such as HapMap. However, data within HapMap generated from the Yoruba population group in Nigeria can tentatively be used as the most representative proxy of a South African sample, to select tagSNPs. This select group of SNPs can complement a selection of SNPs selected based on function, to provide adequate coverage of a candidate gene for an association study.

4.1.4 FGFR1

The computational approach employed here identified *FGFR1* as the top-ranked candidate gene for FASD. Fibroblast growth factors (FGFs) are small polypeptide growth factors, representing a family of at least 22 members with shared structural characteristics. FGFs mediate their biological effects in target cells by binding to and activating a family of four receptor tyrosine kinases – the FGF receptors (FGFR) (Powers et al., 2000). This sets in motion a downstream signalling cascade essential to many biological processes during embryonic development, including cell migration, -survival, -proliferation and -differentiation, among others. The *FGFR* gene family consists of five related genes – *FGFR1*, *FGFR2*, *FGFR3*, *FGFR4* and *FGFR5* which share between 55-72% homology at the protein level (Groth and Lardelli, 2002; Johnson et al., 1990). *FGFR1* was localized to 8p12-p11.2 by Ruta et al. (1988) using in situ hybridization.

- **Protein structure**

The protein encoded by the *FGFR1* gene, as well as the other members of the FGFR family, is evolutionary highly conserved. FGFR family members differ from one another in their ligand affinities and tissue distribution (Groth and Lardelli, 2002). A full-length representative protein of FGFR1 (as shown in Figure 4.1) consists of an extracellular region (corresponding to the 5' end of the *FGFR1* gene), a single hydrophobic membrane-spanning segment and an intracellular tyrosine kinase domain. The extracellular region is composed of a signal peptide followed by three immunoglobulin-like (Ig-like) domains. An acidic box domain is present between the first and second Ig-like domains. The intracellular region consists of two tyrosine kinase domains separated by a 14 residue non-catalytic inter-kinase domain, and is terminated by a C-terminal tail.

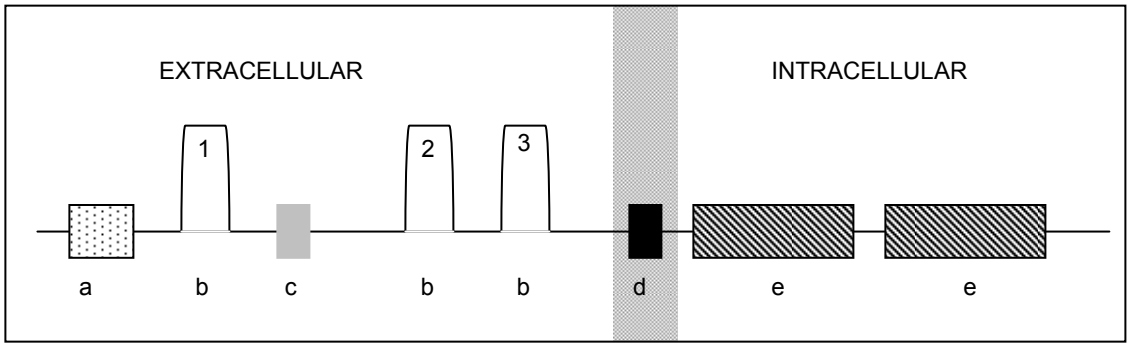

Figure 4.1: Human FGFR1 protein structure. The figure represents a full-length version of the FGFR1 protein (several splice-variants exist). The protein consists of an extracellular and intracellular region, intervened by a transmembrane domain (shown as the shaded area). The extracellular region contains the (a) signal peptide, (b) three Ig-like domains and an (c) acidic box domain. The extracellular region encompasses two tyrosine kinase domains separated by a non-catalytic inter-kinase domain. Modified from Powers et al. (2000) and Groth and Lardelli (2002).

- **Genomic structure and splice variants**

The human *FGFR1* gene is comprised of 19 exons, but alternative splicing results in a diversity of isoforms, expressed in a tissue- and cell-specific manner (Groth and Lardelli, 2002). Several alternatively spliced variants have been described – the Uniprot database currently (December 2007) lists 18 named isoforms for FGFR1 (http://www.ebi.uniprot.org/entry/P11362-15) – though not all variants have been fully characterized (Groth and Lardelli, 2002). Alternative exon usage results in mRNA isoforms that translate into proteins which may be prematurely truncated, or that lack Ig-like or kinase domains. Furthermore, different exon usage can lead to the utilization of different

coding regions for the same Ig-like domains. Figure 4.2 is a general depiction of the longest transcript of *FGFR1*.

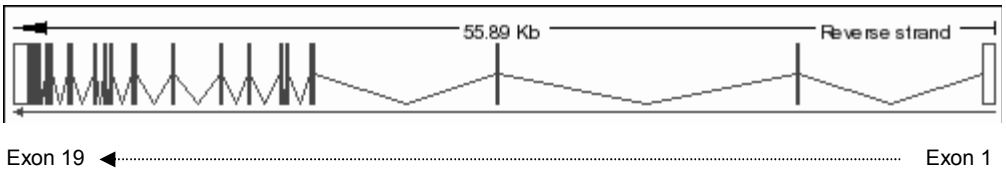

Figure 4.2: The genomic structure of *FGFR1*. The clear boxes indicate untranslated exons. Obtained from http://www.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000077782

The large number of different splice variants isolated from various tissues strongly suggests that *FGFR1* isoforms have specific functions during development and in adult homeostasis. It should however be noted that the number of different *FGFR1* isoforms may possibly be over-estimated. Many *FGFR1* variants have been identified as cDNAs isolated from immortalized cell cultures, implying that although they may well be present in pathological cell state processes, they may not be relevant in the context of normal, healthy tissues. Additionally, the central technique used to identify isoforms, RT-PCR, is highly sensitive and may amplify cDNAs that are mere byproducts of normal *FGFR1* RNA splicing and are present at concentrations too low to have any physiological effect.

- **Disease associations to *FGFR1* mutations**

Mutations in the *FGFR1* gene play a particularly important role in disorders of limb patterning and craniofacial development. Pfeiffer syndrome (OMIM: 101600), which is characterized by craniosynostosis, deviated and enlarged thumbs and big toes, and partial syndactyly of the hands and feet, is associated with a P252R mutation in *FGFR1*. Loss of function mutations in *FGFR1* are the cause of Kallmann syndrome type 2 (KAL2) (OMIM: 147950). In some cases of KAL2, cleft lip and palate and anosmia are present. Defects in *FGFR1* have also been linked to isolated hypogonadotropic hypogonadism (OMIM: 146110), osteoglophonic dysplasia (OMIM: 166250) and non-syndromic trigonocephaly (OMIM: 190440). Chromosomal translocations involving *FGFR1* may also be a cause of stem cell myeloproliferative disorder and stem cell leukemia lymphoma syndrome. The link between craniofacial dysmorphology and *FGFR1*, and the characteristic facial morphology observed in individuals with FASD, underlines the possible link between *FGFR1* and FASD.

- **Proposed involvement in FASD**

The signalling effects caused by FGF stimulation result in, among others, the activation of the MAPK signalling pathway. Recent studies have investigated the effect of controlling second-messenger signalling on neuronal migration in a mouse model of FAS (Kumada et al., 2006). It was shown that experimental manipulation of these second-messenger pathways, through stimulating calcium- and cGMP signalling or inhibiting cAMP signalling, completely reversed the action of ethanol on neuronal migration in vitro as well as in vivo. Each investigated second messenger had multiple but distinct downstream targets, including MAPK.

Few studies have been performed to elucidate the direct link between FGFR1 and alcohol's teratogenic effect during in utero development. One such study investigated the mechanisms involved in the inhibition of functioning of the L1 cell adhesion molecule by ethanol. Ethanol has been shown to inhibit L1-mediated neurite outgrowth of rat cerebellar granule cells (CGN), but it is unclear whether this activation occurs via activation of FGFR1. This study concluded that ethanol disrupted the signalling pathway between L1 clustering and ERK1/2 activation, and that this occurs independently of the FGFR1 pathway in cerebellar granule cells. However, FGFR1 has many cellular downstream effects, and further investigation to elucidate true representation is needed.

The biological outcome of signals generated at the cell surface in response to ligand induced FGFR1 activation is strongly dependent on the cellular context. For example, during early embryo development, FGFR1 plays an important role in control of cell migration, a process crucial for mesodermal patterning and gastrulation, while the activation of FGFR1 signalling in fibroblasts promotes cell proliferation (Schlessinger, 2000). This suggests that common intracellular signalling pathways activated by FGFR1 are able to interact with cell-type specific effector proteins and transcription factors, leading to a specific biological response. The role of FGFR1 in in utero development is therefore evident, making it a valid candidate for FASD.

4.2 SUMMARY OF AIMS

The aim of this analysis was to select SNPs within *FGFR1* based on functional impact for a subsequent case-control association study to evaluate the link between genetic variation and FASD risk. In addition, data from the HapMap project was used to identify tag SNPs for *FGFR1* based on LD blocks observed in the Yoruba population. Furthermore, de novo SNP detection within the regulatory region of the candidate genes was performed. The identification of novel SNPs will augment existing SNP data, which is an important consideration when using a population for which insufficient variation data exist. Based on this, the objectives were:

- To identify novel variation within the regulatory region of the candidate gene *FGFR1*.
- To evaluate the putative functional impact of novel and known SNPs in *FGFR1*
- To identify tag SNPs in *FGFR1*
- To select a subset of SNPs based on function from novel variants and known SNPs within *FGFR1*, and complement these with select tag SNPs for a prospective association study.

4.3 METHODS

All reagent and equipment suppliers are listed in *Addendum C*.

4.3.1 Sample collection

DNA samples for this study were obtained from individuals residing in the Northern Cape towns of De Aar and Upington. Ethics approval for sample collection was obtained from the University of the Witwatersrand, Committee for Research on Human Subjects – Medical (M03-10-20; shown in *Addendum D*). Furthermore, approval has been obtained for experimental use of these samples (M05-01-12; shown in *Addendum D*). The above-mentioned approval was obtained based on informed consent being obtained from all participants, and that a detailed information sheet and oral information session be given before sample collection in the participant's language of choice.

All participants were of mixed-ancestry (also known as the Coloured population), but the geographical distance between the two towns (and therefore the differing parental populations) suggest that the two populations may differ genetically. Therefore, 10

randomly selected control participants from both De Aar and Upington were used for analysis.

4.3.2 DNA extraction and quantification

Samples collected were frozen upon arrival at the laboratory, and batch extractions performed. Batches of samples were defrosted and DNA extracted using the Flexigene DNA kit. The protocol can be described as follows:

Using aseptic techniques, 20 ml of lysis buffer (buffer FG1) was added to a 50 ml polypropelene tube containing 10 ml of whole blood, and mixed by inversion. Samples were centrifuged at 2000 x g for 5 min whereafter the supernatant was discarded. The resultant pellet of white blood cells was resuspended in the denaturation buffer containing protease (buffer FG2) and vortexed until completely homogenized. Samples were incubated for 10 min at 65˚C to aid protein degradation. DNA was precipitated by the addition of 100% isopropanol and centrifugation at 2000 x g for 5 min. The supernatant was discarded, and the DNA pellet washed using 70% ethanol followed by another centrifugation step (2000 x g for 5 min). Hereafter the pellet was air-dried and DNA samples resuspended in 1 ml of hydration buffer (buffer FG3).

DNA was quantified using the Nanodrop® ND-1000 Spectrophotometer. Samples were diluted to a final concentration of 100 ng/µl.

4.3.3 Identification of novel variation within the regulatory region of *FGFR1*

- **Primer design**

The region upstream from the transcription start site and the 5'UTR closest to the translation start site of *FGFR1* were screened for DNA sequence variation, since this region could contain putative regulatory sequences, such as TFBS. The genomic sequence for *FGFR1* gene was obtained from the University of California, Santa Cruz (UCSC) genome database, March 2006 assembly (Kent et al., 2002). Primers were designed using the Primer 3 online tool (Rozen and Skaletsky, 2000), and specificity of the designed primers was determined using the UCSC in-silico PCR tool (Kent et al., 2002), and BLAST analysis (Altschul et al., 1990). Table 4.1 gives a summary of the primers used.

Table 4.1: Summary of PCR primers

PCR Set*	Size	Region*	Forward primer (5'-3')	Reverse primer (5'-3')	T_A
1	500 bp	+56 →-444	ttgccttagcctccgaagta	cagagtgtggctgtgaccag	64
2	586 bp	-11349 → -11935	ggggaagcattttagccact	tacagcctggtctcctttgg	58
3	582 bp	-11843 → -12425	gggggatctcattcagtattcaa	gagacgtgtggttggtttgg	55

* +1 = 'A' of translation start site (ATG). Numbers are based on the genomic sequence

- **PCR and agarose gel electrophoresis**

A PCR temperature gradient was performed for each primer pair to determine the optimal annealing temperature for that specific primer pair. The standard PCR reaction mixture and reaction cycles as mentioned below were used, and a range of temperatures was tested, based on the primer pairs' melting temperature. The PCR gradient was performed in the Eppendorf® Mastercycler® Gradient. Table 4.1 specifies the selected annealing temperatures (T_A) for each reaction.

The PCR reaction mixture (25 μl) consisted of 100ng DNA, 0.2 μM of the forward and reverse primers, 1.25 μM dNTP mix, consisting of equimolar amounts of dATP, dGTP, dCTP and dTTP, 2 μM $MgCl_2$, 0.2 U of Amplitaq Gold Taq polymerase, 1x Amplitaq Gold PCR buffer and deionised H_2O. PCR amplification was performed using either the Eppendorf® Mastercycler® Gradient or the Geneamp® PCR system 2720. The following cycling parameters were used: initial denaturation at 94˚C for 2 min, followed by 45 cycles of denaturation at 94˚C for 30 s, annealing at the determined annealing temperature (Table 4.1) for 15 s and extension at 72˚C for 30 s. The PCR program was completed by a final extension step of 72˚C for 5 min.

Standard electrophoretic techniques were used to resolve the amplified products (Anonymous 2001). A 3% (w/v) agarose gel containing 1mg/L ethidium bromide was used to visualize the amplified products at 6V/cm in 1xTBE buffer (89 mM Tris, 89 mM Boric Acid, 20 mM EDTA, pH 8). A 1 Kb plus DNA molecular weight marker was used as a size standard. Figure 4.3 illustrates the amplification products.

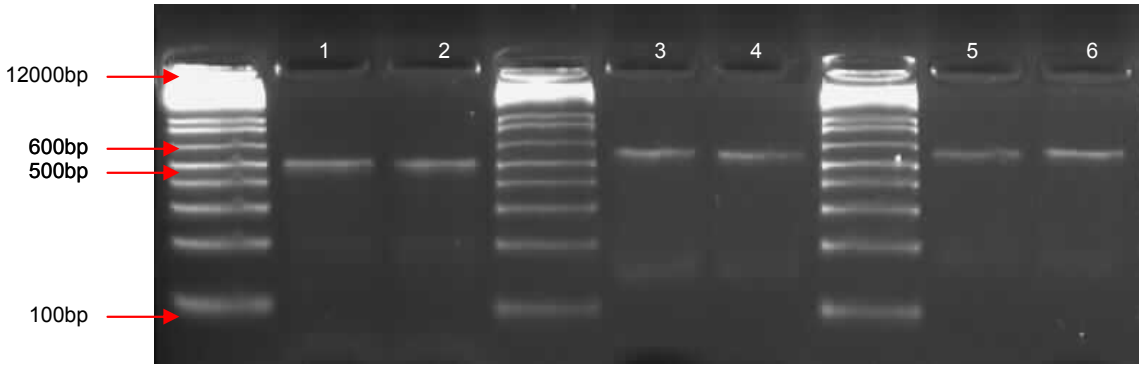

Figure 4.3: Gel electrophoresis of PCR products for the three *FGFR1* regions amplified for subsequent de novo SNP detection. The PCR sets are shown as follow: Lane 1-2 = PCR1; Lane 3-4 = PCR2; Lane 5-6 = PCR3. A 1Kb plus molecular weight marker is shown.

- **DNA sequencing**

DNA sequencing was performed on the ABI 3130 genetic analyser. PCR-amplified products of all 20 samples were cleaned of excess unincorporated PCR reagents using the MultiScreen® $PCR_{\mu 96}$ plate on the Millipore MilliVac® Maxi vacuum manifold. Hereafter the purified PCR samples were prepared for sequencing using the BigDye® Terminator v3.1 cycle sequencing kit. Two reactions per sample were performed – one each using the forward and reverse primer respectively. Each reaction (10 μl) contained 2 μl of the cleaned PCR product, 3 μM of the forward or reverse primer, 1.5 μl of the 5x sequencing buffer and 1 μl of the BigDye Terminator Ready Reaction mix. The cycling parameters were as follow: 25 cycles of: 96˚C denaturation for 30 sec, 50˚C annealing for 15 sec and 60˚C polymerisation for 4 min. The Montage™ 96 well sequencing clean-up system was used to remove impurities.

- **Sequence analysis**

Electropherograms generated by the ABI 3130 genetic analyser were analysed to identify variation using the SeqMan tool from the Lasergene v7 software package (DNASTAR, WI United States).

- **Evaluation of putative functional impact of novel SNPs**

The evaluation of the potential functional impact of novel polymorphisms found by direct sequencing was performed using TFSearch through the FASTSNP module for novel polymorphism evaluation (Yuan et al., 2006). In addition, the program UTRscan (Pesole and Luini, 1999) was used, which allows the searching of user-submitted sequences for any of the patterns within the collection of functional sequence patterns located in 5' and 3' UTR sequences contained in the database UTRsite.

4.3.4 Investigation of functional impact of known SNPs in *FGFR1*

The known SNPs for *FGFR1* were obtained from dbSNP (Sherry et al., 2001). Figure 4.4 shows that intronic SNPs were the most abundant type of SNP within *FGFR1*. In addition, frequency data for selected SNPs were obtained from dbSNP.

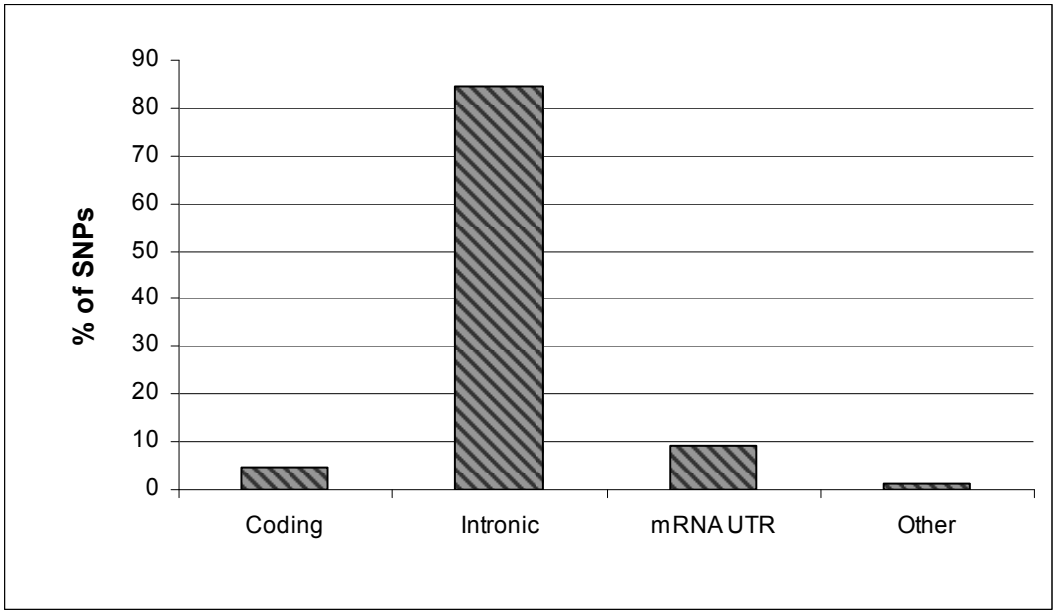

Figure 4.4: Distribution of SNPs in *FGFR1*

The functional impact of known SNPs in *FGFR1* was evaluated in order to prioritize a subset of functional SNPs for association analysis. For most of the analyses (unless otherwise specified) the FASTSNP tool was used (Yuan et al., 2006). FASTSNP is based on an extension of the SNP prioritization strategy proposed by Tabor et al. (Tabor et al., 2002) that entails the calculation of a relative risk for each SNP based on its location and

functional impact. FASTSNP evaluates the functional impact of each SNP by accessing SNP-related information from a variety of online databases, and performing analyses using online tools. Consequently a decision tree is employed to assess the relative risk of each SNP (Figure 4.5). Each risk score has an upper and lower risk ranking, based on the type of impact associated with the SNP.

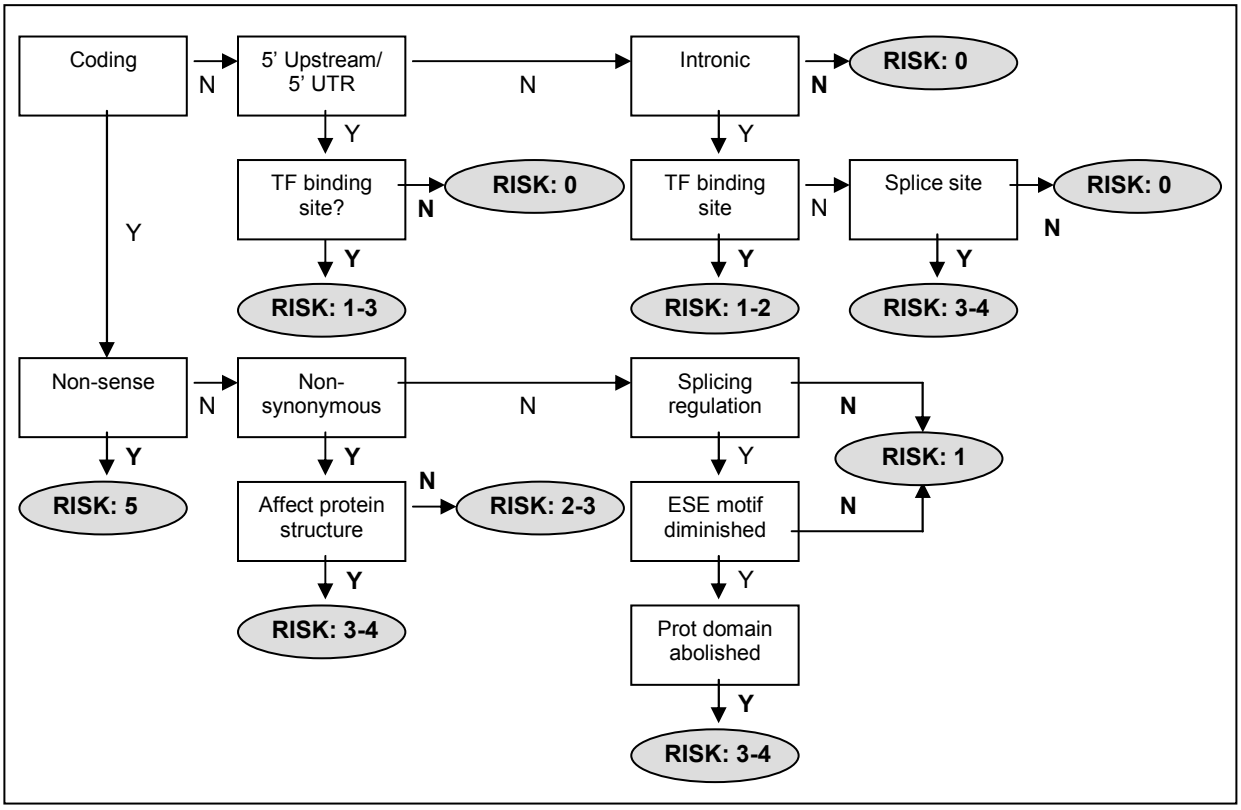

Figure 4.5: The FASTSNP decision tree for prioritizing SNPs based on function. Decision points are indicated by the boxes and the ovals represent terminal points with the risk and class assignments. Y=YES; N=NO. Modified from Yuan et al. (2006). .

The data and tools needed to evaluate functional impact is accessed using the FASTSNP web wrapper agent – i.e. a computational script that allows accession of remote websites, extraction of data and performance of necessary interactions, all from a set location (the FASTSNP website). The added advantage of using a web wrapper agent is that the information used is always current, as there is no need to install the information and the tools needed for analysis locally. SNPs were evaluated for functional impact on the following premises:

- **SNPs affecting protein structure**

The impact of genetic variants on protein structure and function can be predicted by analysis of multiple sequence alignments and evaluation of protein 3D structure. Polyphen exclusively accesses the SWALL database – a comprehensive protein sequence database that combines the high quality of annotation in Swiss-Prot with the completeness of the weekly updated translation of all protein coding sequences from the EMBL nucleotide sequence database. Polyphen assesses whether a non-synonymous SNP results in a change to a residue critical to protein structure by evaluating whether the changes occurs within a disulphide bridge, a binding site, an active site or transmembrane region, among other. The tool further augments the analysis with predictions made with the TMHMM algorithm to predict transmembrane regions (Krogh et al., 2001). The Coils2 program to predict coiled coil regions (Lupas et al., 1991) and the SignalP program to predict signal peptide regions of the protein sequences (Nielsen et al., 1997) are also used in this assessment.

- **SNPs affecting TFBS**

TF Search is an online tool that indicates whether a SNP has a putative regulatory effect, by evaluating if it is positioned within a TFBS, potentially altering the binding of a promoter element (Akiyama, 1998). TFSearch explores highly correlated sequence fragments against the TFMATRIX transcription factor binding site profile database in the TRANSFAC database (Heinemeyer et al., 1998).

- **SNPs affecting exon-intron splicing**

Point mutations frequently cause genetic diseases by disrupting the pattern of pre-mRNA splicing, by inactivating exonic splicing enhancers (ESEs) and resulting in exon skipping. ESEs also appear to be especially important in exons that normally undergo alternative splicing such as *FGFR1*. There are several transcripts of the *FGFR1* gene, which emphasizes the need to investigate SNPs that could influence exonic splice sites. ESEfinder facilitates rapid analysis of exon sequences in order to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40, and SRp55. ESEfinder predicts if a synonymous SNP is located in an exonic splicing enhancer motif, and if the variant would diminish the motif with a different allele (Ahlgren et al., 2002). In addition RescueESS is accessed to serve as a cross-reference and alternative data source for ESEfinder. Finally, FAS-ESS provides identified motifs possessing Exonic Splicing Silencer

(ESS) activity, which is used to predict ESS activity for each SNP allele (Wang et al., 2004).

- **Functional analysis based on homology**

SNP functional evaluation can be performed based on sequence-homology – i.e. assessing whether a SNP position is conserved in model organisms, indicating possible functional importance. One of the earliest tools developed to assess SNP functionality based on sequence homology is the SIFT (sorting intolerant from tolerant) method (Ng and Henikoff, 2003). SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. However, the tool SNPs3D has been shown to perform better than SIFT, and will therefore be used for these analyses (Yue et al., 2006). This online tool uses two support vector machine (SVM) models, based on sequence profiles (Yue and Moult, 2006) and structural stability (Yue et al., 2005). The profile module makes use of the conservation and type of residues observed at a base change position within a protein family to assess function, whereas the structure module analyzes the effect of the resulting amino acid change on protein stability, utilizing structural information. Each SNP is assigned a SVM score, with a negative score indicating that the variant is deleterious to protein function. The SVM is trained on monogenic disease data, so that the definition of deleterious is *'sufficiently damaging to protein function in vivo as to be consistent with a monogenic disease outcome'*.

4.3.5 Selection of tagSNPs from the HapMap database

In addition to the analyses mentioned above, tagSNPs were selected for *FGFR1.* This information will be used to complement the functional SNP selection, if necessary. SNPBrowser v 3.5 was used to select SNPs based on observed LD patterns. SNPs are visualized in their genomic content along with calculated LD maps, putative haplotype blocks and statistical power per gene derived from analysis of more than 3 million SNP genotypes from the full HapMap dataset (De La Vega et al., 2006). SNPbrowser utilized the pair-wise r^2 algorithm to select minimum informative tag-SNPs. Minimum sets of tag SNPs are selected chromosome-wide at three thresholds of pair-wise r^2 through the use of a block-free dynamic programming algorithm framework. The Yoruba population was selected as a representative African population.

4.4 RESULTS

4.4.1 Sequence Analysis

Electropherograms generated by the ABI 3130 genetic analyser were analysed in search of novel polymorphisms within the 5'UTR and promoter region of the *FGFR1* gene in individuals from the mixed ancestry populations of Upington and De Aar. Table 4.2 is a summary of the novel SNPs observed after sequence analysis. Two novel SNPs were only observed in the Upington population, whereas other observed SNPs had similar frequencies in both population groups.

Table 4.2: Novel SNPs observed in the Upington and De Aar mixed ancestry populations.

SNP	Genomic location*	De Aar – Frequency (n=10)				Upington – Frequency (n=10)			
		A	G	T	C	A	G	T	C
g.-12051 G>A	38446172	0.20	0.80	-	-	0.25	0.75	-	-
g.-11798 C>T	38445919	-	-	-	-	-	-	0.05	0.95
g.-11638 C>T	38445759	-	-	-	-	-	-	0.05	0.95
g.-11440 A>G	38445562	0.75	0.25	-	-	0.75	0.25	-	-

**According to UCSC genome database (March 2006 assembly)*

4.4.2 Investigation of functional impact of known SNPs

- **Functional evaluation based on FASTSNP resources**

The functional impact of known SNPs in the *FGFR1* gene was evaluated in order to prioritize a subset of SNPs for association analysis, using the web wrapper agent FASTSNP. As illustration, the decision path taken for evaluation of rs2568231 is portrayed in Figure 4.6. Table 4.3 summarises the SNPs prioritized as having the most deleterious effect. rs17175750, shown in bold in Table 4.3, was also shown to have functional impact by homology analysis (detailed below).

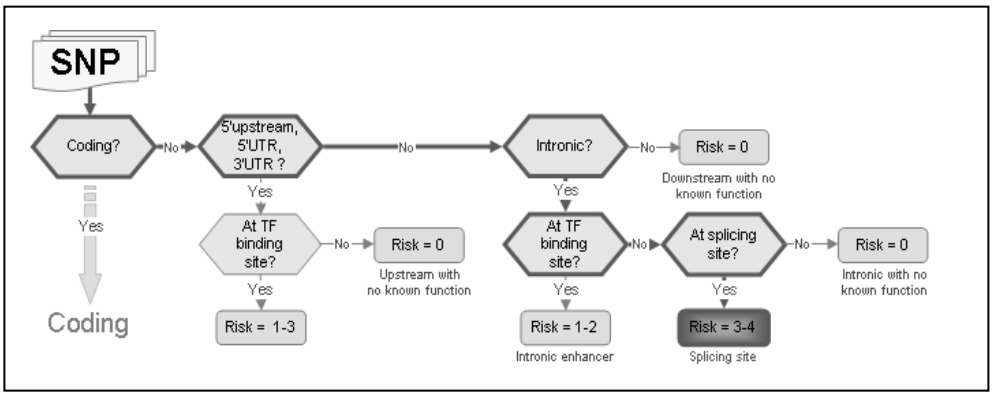

Figure 4.6: An example of FASTSNP output. The evaluation process using the FASTSNP decision tree for SNP rs2568231 is shown**.**

Table 4.3: Known SNPs for the *FGFR1* gene prioritized based on functional impact. Each SNP associated with the gene is given a ranking based on the decision tree used by FASTSNP – with a score of 5 indicating a very high risk, and 0 indicating no risk. Only SNPs evaluated to have some risk are shown.

SNP ID	Putative Functional Effect	RISK	Genomic region	Chromosome position	CDS position	AA Change	AA position
rs2568231	SNP in splicing site	3-4	Intronic	38444649			
rs17175870[1]	Missense/Splicing regulation	2-3	Coding	38406011	1119	T/N	138
rs2915665[1]	synonymous SNP/ Splicing regulation	2-3	Coding	38406370	1051	S	115
rs17175750[1]	Missense/ Splicing regulation	2-3	Coding	38434056	772	R/S	22
rs35494097[1]	synonymous SNP/ Splicing regulation	2-3	Coding	38434068	760	L	18
rs4647908	Regulatory	1-3	5' UTR	38444726	548		
rs4647909	Regulatory	1-3	5' UTR	38444853	421		
rs17182051	Regulatory	1-3	5' Upstream	38445448			
rs328304	Regulatory	1-3	5' Upstream	38445555			
rs17175631	Regulatory	1-3	5' Upstream	38445759			
rs17182030	Regulatory	1-3	5' Upstream	38446159			
rs17182023	Regulatory	1-3	5' Upstream	38446172			
rs7825208	Regulatory	1-3	5' Upstream	38446444			
rs17175617	Regulatory	1-3	5' Upstream	38446637			
rs7829871	Regulatory	1-3	5' Upstream	38446642			
rs17182009	Regulatory	1-3	5' Upstream	38446713			
rs1902171	Regulatory	1-3	5' Upstream	38447599			
rs4082204	Regulatory	1-3	5' Upstream	38448059			
rs12114449	Regulatory	1-3	5' Upstream	38448346			
rs12677355	Regulatory	1-3	5' Upstream	38449100			
rs2604494	Regulatory	1-3	5' Upstream	38450030			

[1]These SNPs were shown to influence splicing regulation by altering ESE sites by ESEFinder.

- **Functional evaluation based on homology**

SNPs3D was used to evaluate the functional impact of the known SNPs for *FGFR1* based on sequence homology. This evaluation revealed that rs17175750 has a putative deleterious effect based on the SVM profile model estimate. The negative SVM score obtained (-1.64) indicates a deleterious effect. Furthermore it was shown that rs17851623, which falls within the 3' UTR of the full transcript of the *FGFR1* gene, had a deleterious effect in many other alternative transcripts of FGFR1.

4.4.3 Evaluation of putative functional impact of novel SNPs

The evaluation of the potential functional impact of novel polymorphisms found by direct sequencing was performed using TFSearch through the FASTSNP module (Yuan et al., 2006) and UTRscan (Pesole and Luini, 1999).

TFSearch was used to evaluate whether a SNP would abolish or insert a putative TFBS in the promoter region of a gene. Table 4.4 shows the results for this exercise. According to this evaluation, SNPs g.-11798 C>T and g.-11440 A>G have no likely functional effect on the regulation of *FGFR1* expression, as neither variant falls within a putative TFBS. UTRscan revealed that none of the novel SNPs fall within the functional sequence patterns known to be present in the 5'UTR of *FGFR1*.

Table 4.4: Evaluation of TFBS affected by novel SNPs found upstream of *FGFR1*. SNPs that have no putative functional effect according to this analysis are shown in italics.

Novel SNP	Genomic location	TFBS present in reference sequence			TFBS present in variant sequence		
		TFMatrix ID	Name	Score[1]	TFMatrix ID	Name	Score[1]
g.-12051 G>A	38446172	M00075	GATA-1	89.8	M00075	GATA-1	86.1
					M00101	CdxA	85.7
g.-11798 C>T	*38445919*	-	-	-	-	-	-
g.-11638 C>T	38445759	M00101	CdxA	87.1	M00240	Nkx-2.	97.7
		M00240	Nkx-2.	86.0	M00101	CdxA	90.0
					M00100	CdxA	89.7
					M00101	CdxA	87.9
					M00101	CdxA	87.1
					M00241	Nkx-2.	85.3
g.-11440 A>G	*38445562*	-	-	-	-	-	-

*[1]Score = 100.0 * (`weighted sum' ± min)/(max ± min) where `max' and `min' are the sum of possible maximum or minimum values of each position of the weighted matrix, respectively. The `weighted sum' is the value calculated by comparing the sequence being evaluated to the weighted matrix. The scoring scheme is a gauge of how well a string matches with the pattern specified by the weighted matrix. The default value for the threshold is 85.0.*

4.4.4 Selection of tagSNPs from the HapMap database

SNPbrowser v3.5 was used to select tagSNPs for *FGFR1.* A total of 10 validated SNPs were available for selection SNPs, of which three failed the minimum allele frequency criterion of MAF > 0.05. Seven SNPs were selected as tagSNPs (Table 4.5), whereas one SNP (rs6983315) was eliminated, as it was tagged by another variant due to LD.

Table 4.5: TagSNPs for *FGFR1* selected by SNPbrowser v3.5

SNP ID	SNPBrowser result
rs2293971	Tagging SNP
rs2915665	Failed MAF* Criterion
rs10108561	Tagging SNP
rs11777067	Failed MAF Criterion
rs6983315	Tagging SNP
rs6987534	Tagging SNP
rs6474354	Tagging SNP
rs10958700	Failed MAF Criterion
rs4733946	Tagging SNP
rs6996321	Tagging SNP

**MAF = Minor Allele Frequency*

4.4.5 Summary

Based on the analyses discussed above, a preliminary subset of SNPs can be selected for a prospective association study. The SNPs selected based on function are listed in Table 4.6. Where available, frequency data for the SNPs are provided. In addition, tagSNPs were selected for *FGFR1*, as shown in Table 4.6. These SNPs can be used in conjunction with the SNPs selected based on functional impact to comprehensively cover the candidate gene.

Table 4.6: A summary of the subset of SNPs in *FGFR1* selected based on putative functional impact. These SNPs will form the foundation for a prospective association study.

SNP	Functional effect
KNOWN SNPs	
rs2568231	SNP in splicing site
rs17175870*	Missense/Splicing regulation
rs2915665*	synonymous SNP/ Splicing regulation
rs17175750*	Missense/Splicing regulation
rs35494097*	synonymous SNP/Splicing regulation
NOVEL SNPs	
g.-12051 G>A	Abolishes TFBS/Creates TFBS
g.-11638 C>T	Abolishes TFBS/Creates TFBS

4.5 DISCUSSION

The rationale of this exercise was to select a subset of SNPs for association analysis based primarily on the functional impact of the genetic variants. All known SNPs within the candidate gene *FGFR1* were computationally evaluated to determine the putative functional impact of the SNP on the gene product's functioning and expression. In addition, tag SNPs were selected to complement the selection made based on potential functional impact.

Table 4.6 summarises the SNPs that will be included in a case-control association study to evaluate the association with risk of FASD development. Only the SNPs that received the highest risk evaluation for functional impact (an evaluation score with a minimum lower risk ranking of 2 on FASTSNP) will initially be included in the association study. In addition to these SNPs, two novel SNPs observed in both the De Aar and Upington populations that were shown to have a putative functional impact will be included in the initial association investigation.

This evaluation was preliminary, and further investigation would be needed to make a final assessment of the most appropriate SNPs to select for an association study. It may be necessary to further sequence sections of *FGFR1* in search of novel SNPs unique to the mixed-ancestry population of South Africa, and to assess the frequency of known SNPs in this population. It should also be noted that other SNPs not included in this selection, may still be associated with the phenotype, and their involvement in FASD can not be excluded at this point. However, this effort illustrates that the functional impact of SNPs on gene expression and protein function can effectively be analysed using available bioinformatics resources.

Case-control tests for association are potentially a powerful strategy for identifying the loci that contribute to complex diseases (Risch and Merikangas, 1996). Case-control association studies involve the typing of genetic polymorphisms in individuals affected by the disease in question (i.e. cases), and a group of healthy, unrelated controls, and comparing compare allele and genotype frequencies to assess the contribution of genetic variants to phenotypes. Association refers to the co-occurrence of an allele, genotype or haplotype with a disease trait, more (or less) frequently than can be readily explained by chance (Mathew, 2001). It is not necessary to genotype all SNPs associated with a candidate gene, and therefore a subset of SNPs is usually selected based either on function or LD and haplotype blocks.

A major pitfall of case-control association studies is that population substructure can result in invalid results being obtained with this approach. Population substructure refers to the presence of different ethnic sub-groups within a population, who differ systematically across loci in their allele frequencies. This phenomenon can contribute to false results due to the fact that both disease frequencies and allele frequencies can differ among subpopulations, therefore resulting in a high rate of significant associations even at markers that are unlinked to any disease locus (Freedman et al., 2004).

Population substructure is a common feature of admixed populations, of which the current study population is an example. The Coloured populations of De Aar and Upington are essentially an admixed group with genetic input from European and various African populations. There are essential differences in the LD and haplotype blocks observed for these parental populations – it has been shown that African populations usually have smaller LD blocks and higher haplotype diversity than European populations (Patterson et al., 2004; Reich et al., 2001). Therefore admixed populations such as the Coloured population groups of De Aar and Upington would be expected to have unique LD patterns, not described in public databases such as the HapMap project database. There are many ways to counter the bias introduced by population substructure, of which one is to consider the frequencies of the genetic variants under investigation within the parental populations. A pitfall of this approach is that it is not always apparent which parental populations contributed to an admixed population, as within admixed groups, individuals may differ widely in the proportion of their ancestry from different sources. This study therefore centred on selecting SNPs for a case-control association based on function, and the selection of novel functional variation within the study populations, rather than basing the selection on haplotype blocks.

This evaluation based on function allowed focusing on a subset of SNPs for an initial association investigation. The methods that were included in this study to evaluate the putative functional impact of SNPs were representative of several aspects, such as the impact of genetic variants on protein structure and function, and putative regulatory effects.

4.6 CONCLUSION

FGFR1 was investigated in this study by evaluating the influence of functional SNPs through computational methods. This computational approach set out here will also be employed to prioritize SNPs based on function from other prioritized candidate genes, specifically those that are involved in the three pathways (the MAPK-, TGF-β and Hedgehog signalling pathways) shown to be significantly over-represented among the top-ranked genes. Using SNPs with possible functional impact in a prospective association study, contributes to the strengthening of the study design for a complex disease gene discovery experiment. Once a statistically significant association is observed for a genetic variant, the information regarding putative impact will lend support for an explanation of the functional role of the polymorphism in question – a prerequisite for result verification. In conclusion, it is proposed that the computational evaluation of genetic variants' impact on function offers a cost-effectiv approach to select a subset of SNPs for future association studies.

Chapter 5

Concluding Remarks

5.1 RATIONALE OF THESIS

This body of work focused on an appropriate study design to computationally predict the genetic contributors to complex diseases. The main points were:

- To develop a method for effective, innovative candidate gene selection and to evaluate the appropriateness and validity of this approach. The advantages of using a computational approach to achieve this aim were explored.
- To select a subset of SNPs from the identified candidate genes for experimental evaluation of association. This selection was made based on the potential functional impact of the genetic variant on protein function and gene expression.

The computational method devised for candidate gene selection was applied to select candidates for FASD, a complex group of syndromes that is primarily caused by in utero alcohol exposure. A putative role for genetic susceptibility to FASD is primarily supported by the observation that FASD does not occur in all children exposed to alcohol during the prenatal period (Bonthius et al., 2004), as well as persuasive lines of evidence from twin concordance studies (Streissguth and Dehaene, 1993) and animal model systems (Sulik, 2005; Sulik et al., 1981). The fact that no genome-wide linkage or association studies have been performed for FASD and that few candidate gene studies have been performed (with conflicting results), was the motivation to use FASD to model the computational approach.

5.2 SUMMARY OF FINDINGS

The computational prioritization method devised to rank genes from a candidate gene list for empirical investigation was based on a simulation of one possible way in which a researcher could select candidate genes i.e. by filtering through various data sources and selecting genes that exhibit the biological characteristics expected to presume a link to the particular disease. A group of 87 genes was prioritized as candidates highlighting many strong candidates from the TGF-β, MAPK and Hedgehog signalling pathways, which are all integral to fetal development and potential targets for alcohol's teratogenic effect. *FGFR1*, a gene that plays a key role in the MAPK signalling pathway, received the highest ranking as a candidate gene (Lombard et al, 2007).

Evaluation of the functional enrichment within the list of top-ranked genes revealed that many of the genes in this ranked list were members of the MAPK-, TGF-β and Hedgehog signalling pathways. These pathways are all integral to embryogenesis and development, and have a potential role in the distinct characteristics associated with FASD, i.e. CNS dysfunction, craniofacial abnormalities and growth retardation. Furthermore it was shown that the prioritized candidates had common promoter elements, implying concurrent regulation and control.

To assess the effectiveness and relative value of this computational approach, X-linked mental retardation (XLMR) was used as a test disease. This implementation resulted in a prioritized gene list with a noted enrichment of known XLMR genes among the top-ranked genes. Furthermore, the top-ranked list contained genes that were biologically relevant to XLMR, and could potentially be yet unknown candidate genes for XLMR. Indeed, many of the top-ranked genes mapped to XLMR candidate regions, confirming their status as good candidates.

Finally, a subset of seven known and novel SNPs were selected within *FGFR1* based on functional impact on gene expression and protein function. Data from the HapMap project were used to identify tag SNPs for *FGFR1* to complement the selection made based on function.

5.3 IMPLICATIONS OF FINDINGS

The results obtained in this study suggest that making a clinically-informed selection from the evidence obtained from literature- and database-mining is an effective approach for candidate disease gene selection and -prioritization. The findings suggest that the method can effectively rank a candidate gene list, resulting in a prioritized selection of candidates that are enriched with biologically plausible candidate genes within pathways related to the disease of interest.

Applying this technique to XLMR showed that, in theory, the method does exhibit the potential to be applied to other complex diseases for candidate gene selection.

A key benefit of this approach is its flexibility – the candidate gene list is prioritized by the binary filtering method without discarding any part of the list. This gives the user the flexibility

to interpret the ranked list in a way appropriate to their future applications. All top-ranked genes may be investigated, or the top-ranked candidates may be examined to identify whether a significant number of the constituents fall within a functionally important pathway, consequently focusing future studies on this select group of genes.

The subset of SNPs selected from the variants within the top-ranked candidate gene, *FGFR1*, would be the focus of a prospective candidate-gene association study investigating the genetic risk involved in FASD development. The importance of de novo SNP discovery in unique, understudied populations such as the mixed-ancestry (Coloured) population of South Africa was also highlighted.

5.4 LIMITATIONS OF CURRENT STUDY

The main pitfall of this method is that it primarily relies on gene annotation to prioritize candidate genes. It is therefore feasible that the method has a biased probability of prioritizing better annotated genes over those that are not as well annotated, regardless of whether these genes are relevant to the disease of interest. The effectiveness of this approach depends on the assumption that the genes under investigation have detailed annotations that clearly define their molecular and physiological functions, in order to avoid erroneous associations.

A further point of consideration is that some clinical understanding of the disease aetiology is needed to aid the binary evaluation, and this process could be partly subjective and researcher-specific. The complexity of multifactorial disorders do not only lie in the intricacy of the genetic contribution, but often also in the fact that much clinical heterogeneity exist. Clinical heterogeneity further complicates the effective application of the described prioritization tool, as this could lead to confounding results.

Employing this approach in selecting candidate genes for a developmental disorder such as FASD, is hampered by the inadequate knowledge available regarding developmental mechanisms. This raises the issue that a further limitation of this technique is that certain disorders will be more amenable to identification of candidates because their etiology is more clearly understood. The developing organism undergoes many rounds of pattern formation, generating complexity with each ensuing round of cell division and with cell differentiation. However, even though the pathways identified using this technique are general fundamental

role players in embryogenesis and development, the technique allowed the focus to fall on specific candidate genes within these pathways for investigation.

It should be considered that high-throughput technology enabling whole-genome association studies may soon supersede the need for techniques such as the one described in this thesis. Lowering costs for both appropriate equipment and genotyping for such analyses implies that association analysis for complex disease may progressively migrate to this option. However, at present the platforms needed to perform such studies are not yet readily available in some countries and remain particularly expensive, implying that the computational prediction of candidate genes is presently still an attractive option in the study design of complex disease gene mapping.

Finally, a significant shortfall is that the technique has not been automated, making the application of this technique to other diseases problematic. Much computational development will be needed to integrate the various steps of the binary filtering process in such a way that the process is applied appropriately, and forms part of the future developments for this study, as discussed below.

5.5 FUTURE DIRECTIONS

A principal future aim for the computational technique described here is to fully automate the process, and have it publicly available for use. This will involve further computational development for which capacity is currently lacking. There are two main considerations:

- This technique is primarily based on data available in the public domain and it must be accepted that not all information available is correct or current. A possible future direction would be to consider how to curate data to only use the most reliable information available. For instance, a date restriction can be set to only include recent (and by default more accurate) information.
- It is expected that automation of this approach will involve a web wrapper agent, to allow always up to date data being available for the evaluation process.

The true test for the appropriateness of the techniques described here would be to establish a statistically significant association between genetic variants and FASD, which can only be

done experimentally. The initial experiment could focus on the subset of SNPs prioritized based on function in *FGFR1*, and other candidate genes.

It has been shown that a case-control based association study is a very powerful approach to assess gene-disease association, and would in this case be an appropriate approach to undertake. Achieving positive and valid results in an association study relies heavily on having an appropriately large sample size to achieve the power and statistical significance to support the evidence for association between a genetic variant and disease. Failing to adhere to the criteria necessary for a valid association study of a complex disease such as FASD would most likely result in a false-positive outcome that would not be reproducible. It has become increasingly apparent that the identification of true genetic association in complex disease will require, among other critical points, sample sizes in the thousands rather than the hundreds (Dahlman et al., 2002). The high profile journal *Nature Genetics* published a list of criteria for association studies to qualify for publication in this journal. These included large sample sizes, small *P* values, and reports of biologically plausible associations. In addition, theses studies should be independently replicated, and the association should be observed both in family-based and population-based studies. Finally, statistical evidence should be highly significant (Nature Genetics Editorial, 2005). This list of criteria also contains the advice that efforts should not be wasted on poorly designed studies and inappropriate statistical analysis. Efforts are therefore underway to collect an appropriately large sample of FASD affected individuals and controls, as well as family members for use in an association study that will test the link between the computationally prioritized candidates and risk of FASD development.

Furthermore, appropriate statistical approaches will be harnessed to overcome the potential pitfalls of using an admixed population for the association study. A series of approaches exist that uses multi-locus genotype data to enable valid case–control tests of association, even in the presence of population structure such as that expected in an admixed population. A technique such as genomic control (detailed in Section 1.1.3) will be employed.

Regarding XLMR, it should be considered that autosomal genes and even structural variation also contribute to mental functioning. Furthermore, the noted allelic- and locus heterogeneity of XLMR, suggest that it is likely that risk factors exist for mental retardation that predispose to, but do not cause the phenotype (Ropers and Hamel, 2005). A future application of the computational process described here to identify and prioritize candidate genes for XLMR, can be expanded to evaluation of the whole-genome for mental retardation candidate genes.

The described method is employed to identify a “most likely” candidate gene list according to known characteristics of the disease. It is an appropriate approach to optimize appropriate candidate gene selection in the face of certain logistical and other restraints associated with the design of a study to identify the genetic risk factors of a complex disease. At best it will highlight a set of biologically plausible candidate gene list most of the time and it is encouraged that it should be used in concert with other candidate gene prediction protocols. Tiffin et al. (2006) recently surveyed some of the methods for computational disease gene identification and concluded that using the methods in concert was more successful in prioritizing candidate genes for disease, than when each was used alone. This review additionally showed that using existing computational methods in concert highlighted potential candidates that are selected by a subset of methods and are missed by the other methods, depending on the type of data examined. This observation gives further evidence that the inclusion of more data sources may positively aid disease gene discovery. The final recommendation is therefore that this technique should be used in conjunction with other computational prediction methods that are appropriate to the specific scenario. Overall it was shown that the genes prioritized using this technique are biologically relevant to the disease, and therefore appropriate for use in a candidate gene association study.

References

Abel, E.L. (1995) An update on incidence of FAS: FAS is not an equal opportunity birth defect. *Neurotoxicol Teratol* **17**, 437-443.

Abidi, F., Hall, B.D., Cadle, R.G., Feldman, G.L., Lubs, H.A., Ouzts, L.V., Arena, J.F., Stevenson, R.E. and Schwartz, C.E. (1999) X-linked mental retardation with variable stature, head circumference, and testicular volume linked to Xq12-q21. *Am J Med Genet* **85**, 223-229.

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**, 55

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* **22**, 773-774.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. and Moreau, Y. (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-544.

Ahlgren, S.C., Thakur, V. and Bronner-Fraser, M. (2002) Sonic hedgehog rescues cranial neural crest from cell death induced by ethanol exposure. *Proc Natl Acad Sci U S A* **99**, 10476-10481.

Ahmad, W., De Fusco, M., ul Haque, M.F., Aridon, P., Sarno, T., Sohail, M., ul Haque, S., Ahmad, M., Ballabio, A., Franco, B. and Casari, G. (1999) Linkage mapping of a new syndromic form of X-linked mental retardation, MRXS7, associated with obesity. *Eur J Hum Genet* **7**, 828-832.

Albrechtsen, A., Castella, S., Andersen, G., Hansen, T., Pedersen, O. and Nielsen, R. (2007) A Bayesian multilocus association method: Allowing for higher-order interaction in association studies. *Genetics* **176**, 1197-1208.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Mol Biol* **215**, 403-410.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-188.

Armfield, K., Nelson, R., Lubs, H.A., Hane, B., Schroer, R.J., Arena, F., Schwartz, C.E. and Stevenson, R.E. (1999) X-linked mental retardation syndrome with short stature, small hands and feet, seizures, cleft palate, and glaucoma is linked to Xq28. *Am J Med Genet* **85**, 236-242.

Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-386.

Aroor, A.R. and Shukla, S.D. (2004) MAP kinase signalling in diverse effects of ethanol. *Life Sci* **74**, 2339-2364.

Astley, S.J. (2004) Diagnostic Guide for Fetal Alcohol Spectrum Disorders: The 4-Digit Diagnostic Code, Seattle: University of Washington

Astley, S.J. and Clarren, S.K. (2000) Diagnosing the full spectrum of fetal alcohol exposed individuals: introducing the 4-digit diagnostic code. *Alcohol Alcohol* **35**, 400-410.

Bailey, J.A. and Eichler, E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**, 552-564.

Bajic, V.B., Choudhary, V. and Hock, C.K. (2004) Content analysis of the core promoter region of human genes. *In Silico Biol* **4**, 109-125.

Bajic, V.B., Tan, S.L., Christoffels, A., Schonbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D.A., Carninci, P. and Hayashizaki, Y. (2006) Mice and men: their promoter properties. *PLoS Genet* **2**, e54.

Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781-791.

Ballabio, A. and Andria, G. (1992) Deletions and translocations involving the distal short arm of the human X chromosome: review and hypotheses. *Hum Mol Genet* **1**, 221-227.

Barnetche, T., Gourraud, P.A. and Cambon-Thomsen, A. (2005) Strategies in analysis of the genetic component of multifactorial diseases; biostatistical aspects. *Transpl Immunol* **14**, 255-266.

Barragan, I., Borrego, S., Abd El-Aziz, M.M., El-Ashry, M.F., Abu-Safieh, L., Bhattacharya, S.S. and Antinolo, G. (2008) Genetic Analysis of FAM46A in Spanish Families with Autosomal Recessive Retinitis Pigmentosa: Characterisation of Novel VNTRs. *Ann Hum Genet* **72**, 26-34.

Barragan, I., Marcos, I., Borrego, S. and Antinolo, G. (2005) Mutation screening of three candidate genes, ELOVL5, SMAP1 and GLULD1 in autosomal recessive retinitis pigmentosa. *Int J Mol Med* **16**, 1163-1167.

Bataller, L., Wade, D.F., Graus, F., Rosenfeld, M.R. and Dalmau, J. (2003) The MAZ protein is an autoantigen of Hodgkin's disease and paraneoplastic cerebellar dysfunction. *Ann Neurol* **53**, 123-127.

Benfante, R., Antonini, R.A., Vaccari, M., Flora, A., Chen, F., Clementi, F. and Fornasari, D. (2005) The expression of the human neuronal alpha3 Na+,K+-ATPase subunit gene is regulated by the activity of the Sp1 and NF-Y transcription factors. *Biochem J* **386**, 63-72.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* **57**, 289-300.

Bertsch, B., Ogden, C.A., Sidhu, K., Le-Niculescu, H., Kuczenski, R. and Niculescu, A.B. (2005) Convergent functional genomics: a Bayesian candidate gene identification approach for complex disorders. *Methods* **37**, 274-279.

Berube, N.G., Jagla, M., Smeenk, C., De Repentigny, Y., Kothary, R. and Picketts, D.J. (2002) Neurodevelopmental defects resulting from ATRX overexpression in transgenic mice. *Hum Mol Genet* **11**, 253-261.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E.W., Wu, B., Doucet, D., Thomas, N.J., Wang, Y., Vollmer, E., Goldmann, T., Seifart, C., Jiang, W., Barker, D.L., Chee, M.S., Floros, J. and Fan, J.B. (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* **16,** 383-393.

Bielawski, D.M., Zaher, F.M., Svinarich, D.M. and Abel, E.L. (2002) Paternal alcohol exposure affects sperm cytosine methyltransferase messenger RNA levels. *Alcohol Clin Exp Res* **26**, 347-51.

Bilotta, J., Barnett, J.A., Hancock, L. and Saszik, S. (2004) Ethanol exposure alters zebrafish development: a novel model of fetal alcohol syndrome. *Neurotoxicol Teratol* **26**, 737-743.

Blake, J.A., Eppig, J.T., Bult, C.J., Kadin, J.A. and Richardson, J.E. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res* **34**, D562-567.

Boehm, S.L. 2nd, Lundahl, K.R., Caldwell, J. and Gilliam, D.M. (1997) Ethanol teratogenesis in the C57BL/6J, DBA/2J, and A/J inbred mouse strains. *Alcohol* **14**, 389-395.

Bonthius, D.J., Karacay, B., Dai, D. and Pantazis, N.J. (2003) FGF-2, NGF and IGF-1, but not BDNF, utilize a nitric oxide pathway to signal neurotrophic and neuroprotective effects against alcohol toxicity in cerebellar granule cell cultures. *Brain Res Dev Brain Res* **140**, 15-28.

Bonthius, D.J., Karacay, B., Dai, D., Hutton, A. and Pantazis, N.J. (2004) The NO-cGMP-PKG pathway plays an essential role in the acquisition of ethanol resistance by cerebellar granule neurons.

Neurotoxicol Teratol **26**, 47-57.

Bortoluzzi, S., Romualdi, C., Bisognin, A. and Danieli, G.A. (2003) Disease genes and intracellular protein networks. *Physiol Genomics* **15**, 223-227.

Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228-237.

Bronner-Fraser, M. (2004) Development. Making sense of the sensory lineage. *Science* **303**, 966-968.

Bruford, E.A., Lush, M.J., Wright, M.W., Sneddon, T.P., Povey, S. and Birney, E. (2007) The HGNC Database in 2008: a resource for the human genome. *Nuc Acid Research* **1-4**,

Buckland, P.R. (2004) Allele-specific gene expression differences in humans. *Hum Mol Genet* **13**, R255-260.

Buckland, P.R. (2006) The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochim Biophys Acta* **1762**, 17-28.

Burgoyne, R.D., O'Callaghan, D.W., Hasdemir, B., Haynes, L.P. and Tepikin, A.V. (2004) Neuronal Ca2+-sensor proteins: multitalented regulators of neuronal function. *Trends Neurosci* **27**, 203-209.

Calella, A.M., Nerlov, C., Lopez, R.G., Sciarretta, C., von Bohlen Und Halbach, O., Bereshchenko, O. and Minichiello, L. (2007) Neurotrophin/Trk receptor signalling mediates C/EBPalpha, -beta and NeuroD recruitment to immediate-early (IE) gene promoters in neuronal cells and requires C/EBPs to induce IE gene transcription. *Neural Develop* **2**, 4.

Carpenter, N.J., Givens, H., Randell, L., Lutz, R. and Miles, J.H. (2000) Clinical characterization and gene mapping of a family with X-linked mental retardation, facial dysmorphism, congenital hip dislocation and skewed pattern of X-inactivation. *Am J Hum Genet* **67**, 1743

Cartegni, L., Chew, S.L. and Krainer, A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**, 285-298.

Cartwright, M.M. and Smith, S.M. (1995) Increased cell death and reduced neural crest cell numbers in ethanol-exposed embryos: partial basis for the fetal alcohol syndrome phenotype. *Alcohol Clin Exp Res* **19**, 378-386.

Castilla, A., Prieto, J. and Fausto, N. (1991) Transforming growth factors beta 1 and alpha in chronic liver disease. Effects of interferon alfa therapy. *N Engl J Med* **324**, 933-940.

Centre for Disease Control (CDC) (2004) Fetal Alcohol Syndrome: Guidelines for Referral and Diagnosis, CDC.

Chai, Y., Ito, Y. and Han, J. (2003) TGF-beta signalling and its functional significance in regulating the fate of cranial neural crest cells. *Crit Rev Oral Biol Med* **14**, 78-88.

Chaudhuri, J.D. (2000) Alcohol and the developing fetus--a review. *Med Sci Monit* **6**, 1031-1041.

Che, Y. and Chen, X. (2004) A multiplexing single nucleotide polymorphism typing method based on restriction-enzyme-mediated single-base extension and capillary electrophoresis. *Anal Biochem* **329**, 220-229.

Chen, D.C., Saarela, J., Nuotio, I., Jokiaho, A., Peltonen, L. and Palotie, A. (2003) Comparison of GenFlex Tag array and Pyrosequencing in SNP genotyping. *J Mol Diagn* **5**, 243-249.

Chen, S.Y., Periasamy, A., Yang, B., Herman, B., Jacobson, K. and Sulik, K.K. (2000) Differential

sensitivity of mouse neural crest cells to ethanol-induced toxicity. *Alcohol* **20**, 75-81.

Chen, Z. and Manley, J.L. (2003) In vivo functional analysis of the histone 3-like TAF9 and a TAF9-related factor, TAF9L. *J Biol Chem* **278**, 35172-35183.

Chernoff, G.F. (1977) The fetal alcohol syndrome in mice: an animal model. *Teratology* **15**, 223-229.

Chudley, A.E., Conry, J., Cook, J.L., Loock, C., Rosales, T. and LeBlanc, N. (2005) Fetal alcohol spectrum disorder: Canadian guidelines for diagnosis. *CMAJ* **172**, S1-21

Chudley, A.E., Tackels, D.C., Lubs, H.A., Arena, J.F., Stoeber, W.P., Kovnats, S., Stevenson, R.E. and Schwartz, C.E. (1999) X-linked mental retardation syndrome with seizures, hypogammaglobulinemia, and progressive gait disturbance is regionally mapped between xq21.33 and Xq23. *Am J Med Genet* **85**, 255-262.

Cilliers, D.D., Parveen, R., Clayton, P., Cairns, S.A., Clarke, S., Shalet, S.M., Black, G.C., Newman, W.G. and Clayton-Smith, J. (2007) A new X-linked mental retardation (XLMR) syndrome with late-onset primary testicular failure, short stature and microcephaly maps to Xq25-q26. *Eur J Med Genet* **50**, 216-223.

Clarren, S.K., Alvord, E.C. Jr, Sumi, S.M., Streissguth, A.P. and Smith, D.W. (1978) Brain malformations related to prenatal exposure to ethanol. *J Pediatr* **92**, 64-67.

Compagni, A., Logan, M., Klein, R. and Adams, R.H. (2003) Control of skeletal patterning by ephrinB1-EphB interactions. *Dev Cell* **5**, 217-230.

Conrad, D.F., Andrews, T.F., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2006) A high-resolution survey of deletion polymorphisms in the human genome. *Nat Genet* **38**, 75-81.

Cordell, H.J. and Clayton, D.G. (2002) A unified stepwise regression approach for evaluating the relative effeects of polymorphisms within a gene using case/control or family data: application to HLA in type I diabetes. *Am J Hum Genet* **70**, 124-141.

Cudd, T.A. (2005) Animal model systems for the study of alcohol teratology. *Exp Biol Med (Maywood)* **230**, 389-393.

Dahlman, I., Eaves, I.A., Kosoy, R., Morrison, V.A., Heward, J., Gough, S.C., Allahabadia, A., Franklyn, J.A., Tuomilehto, J., Tuomilehto-Wolf, E., Cucca, F., Guja, C., Ionescu-Tirgoviste, C., Stevens, H., Carr, P., Nutland, S., McKinney, P., Shield, J.P., Wang, W., Cordell, H.J., Walker, N., Todd, J.A. and Concannon, P. (2002) Parameters for reliable results in genetic association studies in common disease. *Nat Genet* **30**, 149-150.

Daly, A.K. and Day, C.P. (2001) Candidate gene case-control association studies: advantages and potential pitfalls. *Br J Clin Pharmacol* **52**, 489-499.

Damberg, M. (2005) Transcription factor AP-2 and monoaminergic functions in the central nervous system. *J Neural Transm* **112**, 1281-1296.

Das, P. and Golde, T. (2006) Dysfunction of TGF-beta signalling in Alzheimer's disease. *J Clin Invest* **116**, 2855-2857.

Davies, M. (2003) The role of GABAA receptors in mediating the effects of alcohol in the central nervous system. *J Psychiatry Neurosci* **28**, 263-274.

Dawn Teare, M. and Barrett, J.H. (2005) Genetic linkage studies. *Lancet* **366**, 1036-1044.

Day, J.W., Roelofs, R., Leroy, B., Pech, I., Benzow, K. and Ranum, L.P. (1999) Clinical and genetic characteristics of a five-generation family with a novel form of myotonic dystrophy (DM2). *Neuromuscul Disord* **9**, 19-27.

Day, N.L., Zuo, Y., Richardson, G.A., Goldschmidt, L., Larkby, C.A. and Cornelius, M.D. (1999) Prenatal alcohol use and offspring size at 10 years of age. *Alcohol Clin Exp Res* **23**, 863-869.

DeAngelis, J.T., Farrington, W.J. and Tollefsbol, T.O. (2008) An overview of epigenetic assays. *Mol Biotechnol* **38,** 179-183.

de Brouwer, A.P., Yntema, H.G., Kleefstra, T., Lugtenberg, D., Oudakker, A.R., de Vries, B.B., van Bokhoven, H., Van Esch, H., Frints, S.G., Froyen, G., Fryns, J.P., Raynaud, M., Moizard, M.P., Ronce, N., Bensalem, A., Moraine, C., Poirier, K., Castelnau, L., Saillour, Y., Bienvenu, T., Beldjord, C., des Portes, V., Chelly, J., Turner, G., Fullston, T., Gecz, J., Kuss, A.W., Tzschach, A., Jensen, L.R., Lenzner, S., Kalscheuer, V.M., Ropers, H.H. and Hamel, B.C. (2007) Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium. *Hum Mutat* **28**, 207-208.

De La Vega, F.M., Isaac, H.I. and Scafe, C.R. (2006) A tool for selecting SNPs for association studies based on observed linkage disequilibrium patterns. *Pacific Symposium on Biocomputing* **11**, 487-498.

Dearlove, A.M. (2002) High throughput genotyping technologies. *Brief Funct Genomic Proteomic* **1**, 139-150.

Dempfle, A., Wudy, S.A., Saar, K., Hagemann, S., Friedel, S., Scherag, A., Berthold, L.D., Alzen, G., Gortner, L., Blum, W.F., Hinney, A., Nurnberg, P., Schafer, H. and Hebebrand, J. (2006) Evidence for involvement of the vitamin D receptor gene in idiopathic short stature via a genome-wide linkage study and subsequent association studies. *Hum Mol Genet* **15**, 2772-2783.

Dennis, G. Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3.

Dessay, S., Moizard, M.P., Gilardi, J.L., Opitz, J.M., Middleton-Price, H., Pembrey, M., Moraine, C. and Briault, S. (2002) FG syndrome: linkage analysis in two families supporting a new gene localization at Xp22.3. *Am J Med Genet* **112**, 6-11.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**, 997-1004.

Devlin, B., Roeder, K. and Wasserman, L. (2001) Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* **60**, 155-166.

Dick, D.M., Nurnberger, J. Jr, Edenberg, H.J., Goate, A., Crowe, R., Rice, J., Bucholz, K.K., Kramer, J., Schuckit, M.A., Smith, T.L., Porjesz, B., Begleiter, H., Hesselbrock, V. and Foroud, T. (2002) Suggestive linkage on chromosome 1 for a quantitative alcohol-related phenotype. *Alcohol Clin Exp Res* **26** , 1453-1460.

Dimitroulakos, J., Pienkowska, M., Sun, P., Farooq, S., Zielenska, M., Squire, J.A. and Yeger, H. (1999) Identification of a novel zinc finger gene, zf5-3, as a potential mediator of neuroblastoma differentiation. *Int J Cancer* **81**, 97097-97098.

Dlouhy, S.R., Christian, J.C., Haines, J.L., Conneally, P.M. and Hodes, M.E. (1987) Localization of the gene for a syndrome of X-linked skeletal dysplasia and mental retardation to Xq27-qter. *Hum Genet* **75**, 136-139.

Eckhardt, F., Beck, S., Gut, I.G. and Berlin, K. (2004) Future potential of the Human Epigenome Project. *Expert Rev Mol Diagn*, **4**,609-618.

Ehringer, M.A. and Sikela, J.M. (2002) Genomic approaches to the genetics of alcoholism. *Alcohol Res Health* **26**, 181-192.

Eppig, J.T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A. and and the members of the Mouse

Genome Database Group (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* **33**, D471-475.

Esteller, M. (2008) Epigenetics in cancer. *N Engl J Med* **358**, 1148-1159.

Falk, C.T. and Rubinstein, P. (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* **51**, 227-233.

Farber, N.B. and Olney, J.W. (2003) Drugs of abuse that cause developing neurons to commit suicide. *Brain Res Dev Brain Res* **147**, 37-45.

Fendri, K., Kefi, M., Hentati, F. and Amouri, R. (2006) Genetic heterogeneity within a consanguineous family involving the LGMD 2D and the LGMD 2C genes. *Neuromuscul Disord* **16**, 316-320.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A. and Searle, S. (2008) Ensembl 2008. *Nucleic Acids Res* **36**, D707-714.

Franke, L., Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M. and Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* **78**, 1011-1025.

Fredericks, D.W. and Williams, W.L. (1998) New definition of mental retardation for the American Association of Mental Retardation. *Image J Nurs Sch* **30**, 53-56.

Freedman, M.L., Reich, D., Penney, K.L., McDonald, G.J., Mignault, A.A., Patterson, N., Gabriel, S.B., Topol, E.J., Smoller, J.W., Pato, C.N., Pato, M.T., Petryshen, T.L., Kolonel, L.N., Lander, E.S., Sklar, P., Henderson, B., Hirschhorn, J.N. and Altshuler, D. (2004) Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388-393.

Freudenberg, J. and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* **18 Suppl 2**, S110-115.

Galvan, V., Logvinova, A., Sperandio, S., Ichijo, H. and Bredesen, D.E. (2003) Type 1 insulin-like growth factor receptor (IGF-IR) signalling inhibits apoptosis signal-regulating kinase 1 (ASK1). *J Biol Chem* **278**, 13325-13332.

Garrigue-Antar, L., Munoz-Antonia, T., Antonia, S.J., Gesmonde, J., Vellucci, V.F. and Reiss, M. (1995) Missense mutations of the transforming growth factor beta type II receptor in human head and neck squamous carcinoma cells. *Cancer Res* **55**, 3982-3987.

Garro, A.J., McBeth, D.L., Lima, V. and Lieber, C.S. (1991) Ethanol consumption inhibits fetal DNA methylation in mice: implications for the fetal alcohol syndrome. *Alcohol Clin Exp Res* **15**, 395-398.

George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D. and Wouters, M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* **34**, e130

Gilliam, D.M., Mantle, M.A., Barkhausen, D.A. and Tweden, D.R. (1997) Effects of acute prenatal ethanol administration in a reciprocal cross of C57BL/6J and short-sleep mice: maternal effects and nonmaternal factors. *Alcohol Clin Exp Res* **21**, 28-34.

Glaser, R.L., Ramsay, J.P. and Morison, I.M. (2006) The imprinted gene and parent-of-origin effect

database now includes parental origin of de novo mutations. *Nucleic Acids Res* **34**, D29-31.

Goldgar, D.E. (2001) Major strengths and weaknesses of model-free methods. *Adv Genet* **42**, 241-251.

Goldstein, J.A. and Lalwani, A.K. (2002) Further evidence for a third deafness gene within the DFNA2 locus. *Am J Med Genet* **108**, 304-309.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007) The human disease network. *PNAS.***104**, 8685-8690.

Gomes, F.C., Sousa Vde, O. and Romao, L. (2005) Emerging roles for TGF-beta1 in nervous system development. *Int J Dev Neurosci* **23**, 413-424.

Goodlett, C.R. and Horn, K.H. (2001) Mechanisms of alcohol-induced damage to the developing nervous system. *Alcohol Res Health* **25**, 175-184.

Gordi, T. and Khamis, H. (2004) Simple Solution to a Common Statistical Problem: Interpreting Multiple Tests. *Clin Ther* **26**, 780-786.

Graham, C.A., Redmond, R.M. and Nevin, N.C. (1991) X-linked clinical anophthalmos. Localization of the gene to Xq27-Xq28. *Ophthalmic Paediatr Genet* **12**, 43-48.

Greene, L.A., Liu, D.X., Troy, C.M. and Biswas, S.C. (2007) Cell cycle molecules define a pathway required for neuron death in development and disease. *Biochim Biophys Acta* **1772**, 392-401.

Grisel, J.E., Metten, P., Wenger, C.D., Merrill, C.M. and Crabbe, J.C. (2002) Mapping of quantitative trait loci underlying ethanol metabolism in BXD recombinant inbred mouse strains. *Alcohol Clin Exp Res* **26**, 610-616.

Groth, C. and Lardelli, M. (2002) The structure and function of vertebrate fibroblast growth factor receptor 1. *Int J Dev Biol* **46**, 393-400.

Guo, S.W. and Thompson, E.A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y. and et, a.l. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-238.

Hahn, S.A., Schutte, M., Hoque, A.T., Moskaluk, C.A., da Costa, L.T., Rozenblum, E., Weinstein, C.L., Fischer, A., Yeo, C.J., Hruban, R.H. and Kern, S.E. (1996) DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science* **271**, 350-353.

Haines, J.L. and Pericak-Vance, M.A. (2006) Designing a study for identifying genes in complex traits. In: Haines, J.L. and Pericak-Vance, M.A.e., (Eds.) *Genetic analysis of complex disease*, Second edition. pp. 455-467. New Jersey USA: John Wiley & Sons

Han, D.C., Shen, T.L., Miao, H., Wang, B. and Guan, J.L. (2002) EphB1 associates with Grb7 and regulates cell migration. *J Biol Chem* **277**, 45655-45661.

Handoko, H.Y., Nancarrow, D.J., Mowry, B.J. and McGrath, J.J. (2006) Polymorphisms in the vitamin D receptor and their associations with risk of schizophrenia and selected anthropometric measures. *Am J Hum Biol* **18**, 415-417.

Hao, H.N., Parker, G.C., Zhao, J., Barami, K. and Lyman, W.D. (2003) Differential responses of human neural and hematopoietic stem cells to ethanol exposure. *J Hematother Stem Cell Res* **12**, 389-399.

Haravuori, H., Siitonen, H.A., Mahjneh, I., Hackman, P., Lahti, L., Somer, H., Peltonen, L., Kestila, M. and Udd, B. (2004) Linkage to two separate loci in a family with a novel distal myopathy

phenotype (MPD3). *Neuromuscul Disord* **14**, 183-187.

Hardy, K. (1999) Apoptosis in the human embryo. *Rev Reprod* **4**, 125-134.

Hasty, J., McMillen, D., Isaacs, F. and Collins, J.J. (2001) Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet* **2**, 268-279.

Heath, A.C. and Nelson, E.C. (2002) Effects of the interaction between genotype and environment. Research into the genetic epidemiology of alcohol dependence. *Alcohol Res Health* **26**, 193-201.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A., Podkolodny, N.L. and Kolchanov, N.A. (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res* **26**, 362-367.

Hoh, J. and Ott, J. (2004) Genetic dissection of diseases: design and methods. *Curr Opin Genet Dev* **14**, 229-232.

Hoyme, H.E., May, P.A., Kalberg, W.O., Kodituwakku, P., Gossage, J.P., Trujillo, P.M., Buckley, D.G., Miller, J.H., Aragon, A.S., Khaole, N., Viljoen, D.L., Jones, K.L. and Robinson, L.K. (2005) A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: clarification of the 1996 institute of medicine criteria. *Pediatrics* **115**, 39-47.

Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphrey, S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* **74**, 289-298.

Hsiao, S.H., DuBois, D.W., Miranda, R.C. and Frye, G.D. (2004) Critically timed ethanol exposure reduces GABAAR function on septal neurons developing in vivo but not in vitro. *Brain Res* **1008**, 69-80.

Hsu, S.H., Hsieh-Li, H.M., Huang, H.Y., Huang, P.H. and Li, H. (2005) bHLH-zip transcription factor Spz1 mediates mitogen-activated protein kinase cell proliferation, transformation, and tumorigenesis. *Cancer Res* **65**, 4041-4050.

Huang, T.H., Hejtmancik, J.F., Edwards, A., Pettigrew, A.L., Herrera, C.A., Hammond, H.A., Caskey, C.T., Zoghbi, H.Y. and Ledbetter, D.H. (1991) Linkage of the gene for an X-linked mental retardation disorder to a hypervariable (AGAT)n repeat motif within the human hypoxanthine phosphoribosyltransferase (HPRT) locus (Xq26). *Am J Hum Genet* **49**, 1312-1319.

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971-983.

Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Res* **18**, 644-652

Iles, M.M. (2008) What Can Genome-Wide Association Studies Tell Us about the Genetics of Common Disease? *Plos Genet* **4**, e33.

Ikonomidou, C., Bittigau, P., Ishimaru, M.J., Wozniak, D.F., Koch, C., Genz, K., Price, M.T., Stefovska, V., Horster, F., Tenkova, T., Dikranian, K. and Olney, J.W. (2000) Ethanol-induced apoptotic neurodegeneration and fetal alcohol syndrome. *Science* **287**, 1056-1560.

Ingham, P.W. and McMahon, A.P. (2001) Hedgehog signalling in animal development: paradigms and principles. *Genes Dev* **15**, 3059-3087.

International HapMap Consortium (2003) The International HapMap Project. *Nature* **426**, 789-796.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**, 1299-

1320.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861.

Ionita-Laza, I., Perry, G.H., Raby, B.A., Klanderman, B., Lee, C., Laird, N.M., Weiss, S.T. and Lange, C. (2008) On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. *Genet Epidemiol* **32**, 273-284

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. and Contopoulos-Ioannidis, D.G. (2001) Replication validity of genetic association studies. *Nat Genet* **29**, 306-309.

Isaka, Y., Brees, D.K., Ikegaya, K., Kaneda, Y., Imai, E., Noble, N.A. and Border, W.A. (1996) Gene therapy by skeletal muscle expression of decorin prevents fibrotic disease in rat kidney. *Nat Med* **2**, 418-423.

Ishitani, K., Yoshida, T., Kitagawa, H., Ohta, H., Nozawa, S. and Kato, S. (2003) p54nrb acts as a transcriptional coactivator for activation function 1 of the human androgen receptor. *Biochem Biophys Res Commun* **306**, 660-665.

Itoh, F., Ishizaka, Y., Tahira, T., Yamamoto, M., Miya, A., Imai, K., Yachi, A., Takai, S., Sugimura, T. and Nagao, M. (1992) Identification and analysis of the ret proto-oncogene promoter region in neuroblastoma cell lines and medullary thyroid carcinomas from MEN2A patients. *Oncogene* **7**, 1201-1206.

Jacobson, S.W., Carr, L.G., Croxford, J., Sokol, R.J., Li, T.K. and Jacobson, J.L. (2006) Protective effects of the alcohol dehydrogenase-ADH1B allele in children exposed to alcohol during pregnancy. *J Pediatr* **148**, 30-37.

Jakowlew, S.B. (2006) Transforming growth factor-beta in cancer and metastasis. *Cancer Metastasis Rev* **25**, 435-457.

Jamerson, P.A., Wulser, M.J., Kimler, B.F. (2004) Neurobehavioral effects in rat pups whose sires were exposed to alcohol. *Brain Res Dev Brain Res* **149,**103-111.

Jehee, F.S., Rosenberg, C., Krepischi-Santos, A.C., Kok, F., Knijnenburg, J., Froyen, G., Vianna-Morgante, A.M., Opitz, J.M. and Passos-Bueno, M.R. (2005) An Xq22.3 duplication detected by comparative genomic hybridization microarray (arrayCGH) defines a new locus (FGS5) for FG syndrome. *Am J Med Genet* **139A**, 221-226.

Ji, M., Hou, P., Li, S., He, N. and Lu, Z. (2004) Microarray-based method for genotyping of functional single nucleotide polymorphisms using dual-color fluorescence hybridization. *Mutat Res* **548**, 97-105.

Jirtle, R.L. and Skinner, M.K. (2007) Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **8**, 253-262.

Jirtle, R.L., Sander, M. and Barrett, J.C. (2000) Genomic imprinting and environmental disease susceptibility. *Environ Health Perspect* **108**, 271-278.

Johnson, D.E., Lee, P.L., Lu, J. and Williams, L.T. (1990) Diverse forms of a receptor for acidic and basic fibroblast growth factors. *Mol Cell Biol* **10**, 4728-4736.

Johnson, J.P., Nelson, R. and Schwartz, C.E. (1998) A family with mental retardation, variable macrocephaly and macro-orchidism, and linkage to Xq12-q21. *J Med Genet* **35**, 1026-1030.

Jones, K.L. and Smith, D.W. (1973) Recognition of the fetal alcohol syndrome in early infancy. *Lancet* **2**, 999-1001.

Jones, K.L., Smith, D.W., Ulleland, C.N. and Streissguth, P. (1973) Pattern of malformation in offspring

of chronic alcoholic mothers. *Lancet* **1**, 1267-1271.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C.V., McCarthy, M.I., Hide, T. and Hide, W. (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* **13**, 1222-1230.

Kent, W.J., Hsu, F., Karolchik, D., Kuhn, R.M., Clawson, H., Trumbower, H. and Haussler, D. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res* **15**, 737-741.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res* **12**, 996-1006.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. and Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073-1080.

Khalil, N. and Greenberg, A.H. (1991) The role of TGF-beta in pulmonary fibrosis. *Ciba Found Symp* **157**, 194-207; discussion 207-211.

Khaole, N.K. and Li, T.K (2000). Protective alcohol dehydrogenase genotypes for FAS and blood alcohol profiles among mothers of FAS children. Paper presented at *The Annual Meeting of the Research Society on Alcoholism*, June 24–29, 2000, Denver, Colo.

Kim, S. and Misra, A. (2007) SNP genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng* **9**, 289-320.

Kleefstra, T. and Hamel, B.C.J. (2005) X-linked mental retardation: further lumping, splitting and emerging phenotypes. *Clin Genet* **67**, 451-467.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. and Hoh, J. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389.

Knight, J.C. (2005) Regulatory polymorphisms underlying complex disease traits. *J Mol Med* **83**, 97-109.

Knoll, J.H., Nicholls, R.D., Magenis, R.E., Graham, J.M., Lalande, M. and Latt, S.A. (1989) Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *Am J Med Genet* **32**, 285-290.

Kobayashi, A., Yamagiwa, H., Hoshino, H., Muto, A., Sato, K., Morita, M., Hayashi, N., Yamamoto, M. and Igarashi, K. (2000) A combinatorial code for gene expression generated by transcription factor Bach2 and MAZR (MAZ-related factor) through the BTB/POZ domain. *Mol Cell Biol* **20**, 1733-1746.

Koenig, M., Hoffman, E.P., Bertelson, C.J., Monaco, A.P., Feener, C. and Kunkel, L.M. (1987) Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509-517.

Koller, D.L., Peacock, M., Lai, D., Foroud, T. and Econs, M.J. (2004) False positive rates in association studies as a function of degree of stratification. *J Bone Miner Res* **19**, 1291-1295.

Korbel, J.O., Doerks, T., Jensen, L.J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S.D., Andrade, M.A. and Bork, P. (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3**, e134

Krens, S.F., Spaink, H.P. and Snaar-Jagalska, B.E. (2006) Functions of the MAPK family in vertebrate-development. *FEBS Lett* **580**, 4984-4990.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580.

Kruglyak, L. and Nickerson, D.A. (2001) Variation is the spice of life. *Nat Genet* **27**, 234-236.

Kumada, T., Lakshmana, M.K. and Komuro, H. (2006) Reversal of neuronal migration in a mouse model of fetal alcohol syndrome by controlling second-messenger signallings. *J Neurosci* **26**, 742-756.

Kwok, P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet* **2**, 235-258.

Laird, N.M. and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* **7**, 385-394.

Lange-Dohna, C., Zeitschel, U., Gaunitz, F., Perez-Polo, J.R., Bigl, V. and Rossner, S. (2003) Cloning and expression of the rat BACE1 promoter. *J Neurosci Res* **73**, 73-80.

Lee, H.Y., Kleber, M., Hari, L., Brault, V., Suter, U., Taketo, M.M., Kemler, R. and Sommer, L. (2004) Instructive role of Wnt/beta-catenin in sensory fate specification in neural crest stem cells. *Science* **303**, 1020-1023.

Lee, T.Y., Chin, G.S., Kim, W.J., Chau, D., Gittes, G.K. and Longaker, M.T. (1999) Expression of transforming growth factor beta 1, 2, and 3 proteins in keloids. *Ann Plast Surg* **43**, 179-184.

Lehrke, R. (1972) Theory of X-linkage of major intellectual traits. *Am J Ment Defic* **76**, 611-619.

Lezirovitz, K., Pardono, E., de Mello Auricchio, M.T., de Carvalho, E. Silva FL, Lopes, J.J., Abreu-Silva, R.S., Romanos, J., Batissoco, A.C. and Mingroni-Netto, R.C. (2008) Unexpected genetic heterogeneity in a large consanguineous Brazilian pedigree presenting deafness. *Eur J Hum Genet* **16**, 89-96.

Lichten, W. and Simon, E.W. (2007) Defining mental retardation: a matter of life or death. *Intellect Dev Disabil* **45**, 335-46.

Lim, H.N. and van Oudenaarden, A. (2007) A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat Genet* **39**, 269-275.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-182.

Lopez-Bigas, N. and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108-31014.

Lopez-Bigas, N., Blencowe, B.J. and Ouzounis, C.A. (2006) Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics* **22**, 269-277.

Louis, E.J. and Dempster, E.R. (1987) An exact test for Hardy-Weinberg and multiple alleles. *Biometrics* **43**, 805-811.

Lupas, A., Van Dyke, M. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252**, 1162-1164.

Madsen, E. and Gitlin, J.D. (2007) Copper and iron disorders of the brain. *Annu Rev Neurosci* **30**, 317-

337.

*Maier, S.E. and West, J.R. (2001) Drinking patterns and alcohol-related birth defects. Alcohol Res Health **25**, 168-174.*

Manning, M.A. and Hoyme, E.H. (2007) Fetal alcohol spectrum disorders: a practical clinical approach to diagnosis. *Neurosci Biobehav Rev* **31**, 230-238.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W., Vogelstein, B. and et, a.l. (1995) Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268**, 1336-1338.

Martin, D.M., Probst, F.J., Camper, S.A. and Petty, E.M. (2000) Characterisation and genetic mapping of a new X linked deafness syndrome. *J Med Genet* **37**, 836-841.

Mashima, Y., Yamamoto, S., Inoue, Y., Yamada, M., Konishi, M., Watanabe, H., Maeda, N., Shimomura, Y. and Kinoshita, S. (2000) Association of autosomal dominantly inherited corneal dystrophies with BIGH3 gene mutations in Japan. *Am J Ophthalmol* **130**, 516-517.

Masseroli, M., Galati, O., Manzotti, M., Gibert, K. and Pinciroli, F. (2005) Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. *BMC Bioinf* **6**, S18

Masseroli, M., Martucci, D. and Pinciroli, F. (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nuc Acid Res* **32**, 293-300.

Mathew, C. (2001) Science, medicine, and the future: Postgenomic technologies: hunting the genes for common disorders. *BMJ* **322**, 1031-1034.

Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat Rev Genet* **9**, 9-14.

May, P.A., Gossage, J.P., Brooke, L.E., Snell, C.L., Marais, A.S., Hendricks, L.S., Croxford, J.A. and Viljoen, D.L. (2005) Maternal risk factors for fetal alcohol syndrome in the Western cape province of South Africa: a population-based study. *Am J Public Health* **95**, 1190-1199.

May, P.A., Gossage, J.P., Marais, A.S., Adnams, C.M., Hoyme, H.E., Jones, K.L., Robinson, L.K., Khaole, N.C., Snell, C., Kalberg, W.O., Hendricks, L., Brooke, L., Stellavato, C. and Viljoen, D.L. (2007) The epidemiology of fetal alcohol syndrome and partial FAS in a South African community. *Drug Alcohol Depend* **88**, 259-271.

May, P.A., Gossage, J.P., White-Country, M., Goodhart, K., Decoteau, S., Trujillo, P.M., Kalberg, W.O., Viljoen, D.L. and Hoyme, H.E. (2004) Alcohol consumption and other maternal risk factors for fetal alcohol syndrome among three distinct samples of women before, during, and after pregnancy: the risk is relative. *Am J Med Genet C Semin Med Genet* **127**, 10-20.

McAllister, K.A., Grogg, K.M., Johnson, D.W., Gallione, C.J., Baldwin, M.A., Jackson, C.E., Helmbold, E.A., Markel, D.S., McKinnon, W.C., Murrell, J. and et, a.l. (1994) Endoglin, a TGF-beta binding protein of endothelial cells, is the gene for hereditary haemorrhagic telangiectasia type 1. *Nat Genet* **8**, 345-351.

McCarver, D.G., Thomasson, H.R., Martier, S.S., Sokol, R.J. and Li, T. (1997) Alcohol dehydrogenase-2*3 allele protects against alcohol-related birth defects among African Americans. *J Pharmacol Exp Ther* **283**, 1095-1101.

Meloni, I., Bruttini, M., Longo, I., Mari, F., Rizzolio, F., D'Adamo, P., Denvriendt, K., Fryns, J.P., Toniolo, D. and Renieri, A. (2000) A mutation in the rett syndrome gene, MECP2, causes X-linked mental retardation and progressive spasticity in males. *Am J Hum Genet* **67**, 982-985.

Melvin, E.C. and Speer, M.C. (2006) Basic concepts in genetics and linkage analysis. In: *Haines, J.L. and Pericak-Vance, M.A.e., (Eds.) Genetic analysis of complex disease, Second edition. pp. 1-46. New Jersey USA: John Wiley and Sons Inc*

Michelson, P., Hartwig, C., Schachner, M., Gal, A., Veske, A. and Finckh, U. (2002) Missense mutations in the extracellular domain of the human neural cell adhesion molecule L1 reduce neurite outgrowth of murine cerebellar neurons. *Hum Mutat* **20**, 481-482.

Miles, J.H. and Carpenter, N.J. (1991) Unique X-linked mental retardation syndrome with fingertip arches and contractures linked to Xq21.31. *Am J Med Genet* **38**, 215-23.

Miller, M.W. and Luo, J. (2002) Effects of ethanol and transforming growth factor beta (TGF beta) on neuronal proliferation and nCAM expression. *Alcohol Clin Exp Res* **26**, 1281-1285.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S. and Cheung, V.G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747.

Moss, T.J., Wallrath, L.L. (2007) Connections between epigenetic gene silencing and human disease. *Mutat Res* **618**, 163-174.

Nady, N., Min, J., Kareta, M.S., Chédin, F. and Arrowsmith, C.H. (2008) A SPOT on the chromatin landscape? Histone peptide arrays as a tool for epigenetic research. *Trends Biochem Sci* IN PRESS (Early online publication).

Namkung, J., Kim, Y. and Park, T. (2005) Whole-genome association studies of alcoholism with loci linked to schizophrenia susceptibility. *BMC Genet* **6 Suppl 1**, S9.

Nature Genetics Editorial (1999) Freely associating. *Nat Genet* **22**, 1-2.

Nature Genetics Editorial (2005) Framework for a fully powered risk engine. *Nat Genet* **37**, 1153

Ng, P.C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1-6.

Nordfors, L., Jansson, M., Sandberg, G., Lavebratt, C., Sengul, S., Schalling, M. and Arner, P. (2002) Large-scale genotyping of single nucleotide polymorphisms by Pyrosequencingtrade mark and validation against the 5'nuclease (Taqman®) assay. *Hum Mutat* **19**, 395-401.

Nowotny, P., Kwon, J.M. and Goate, A.M. (2001) SNP analysis to dissect human traits. *Curr Opin Neurobiol* **11**, 637-641.

Nyholt, D.R. (2001) Genetic case-control association studies – correcting for multiple testing. *Hum Genet* **109**, 564-565.

O'Connor, T.P. and Crystal, R.G. (2006) Genetic medicines: treatment strategies for hereditary disorders. *Nature Reviews Genetics* **7**, 261-276.

Ogawa, T., Kuwagata, M., Ruiz, J. and Zhou, F.C. (2005) Differential teratogenic effect of alcohol on embryonic development between C57BL/6 and DBA/2 mice: a new view. *Alcohol Clin Exp Res* **29**, 855-863.

Okamoto, S., Sherman, K., Bai, G. and Lipton, S.A. (2002) Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal

differentiation. *Brain Res Mol Brain Res* **107**, 89-96.

Oosterwijk, J.C., Wischmeijer, A., Losekoot, M., Haraldson, A., Theunissen, G. and van Gelderen, I. (1999) A new X-linked mental retarda-tion syndrome with distal limb defects, hearing impairment, verrucosis and immunodeficiency. *Am J Hum Genet* **65**, 1898

Orton, R.J., Sturm, O.E., Vyshemirsky, V., Calder, M., Gilbert, D.R. and Kolch, W. (2005) Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. *Biochem J* **392**, 249-261.

Oti, M. and Brunner, H. (2007) The modular nature of genetic diseases. *Clin Genet* **71**, 1-11.

Pan, H., Zuo, L., Choudhary, V., Zhang, Z., Leow, S.H., Chong, F.T., Huang, Y., Ong, V.W., Mohanty, B., Tan, S.L., Krishnan, S.P. and Bajic, V.B. (2004) Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucleic Acids Res* **32**, W230-234.

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., Daly, M.J. and Reich, D. (2004) Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74**, 979-1000.

Perez-Iratxeta, C., Wjst, M., Bork, P. and Andrade, M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet* **6**, 45

Perrotti, D., Melotti, P., Skorski, T., Casella, I., Peschle, C. and Calabretta, B. (1995) Overexpression of the zinc finger protein MZF1 inhibits hematopoietic development from embryonic stem cells: correlation with negative regulation of CD34 and c-myb promoter activity. *Mol Cell Biol* **15**, 6075-6087.

Perry ,G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., Park, H.S., Kim, J.I., Seo, J.S., Yakhini, Z., Laderman, S., Bruhn, L. and Lee, C. (2008) The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-695

Pesole, G. and Luini, S. (1999) Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. *Trends Genet* **15**, 378.

Piluso, G., Carella, M., D'Avanzo, M., Santinelli, R., Carrano, E.M., D'Avanzo, A., D'Adamo, A.P., Gasparini, P. and Nigro, V. (2003) Genetic heterogeneity of FG syndrome: a fourth locus (FGS4) maps to Xp11.4-p11.3 in an Italian family. *Hum Genet* **112**, 124-130.

Powers, C.J., McLeskey, S.W. and Wellstein, A. (2000) Fibroblast growth factors, their receptors and signalling. *Endocr Relat Cancer* **7**, 165-197.

Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000) Association mapping in structured populations. *Am J Hum Genet* **67**, 170-181.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez. J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R. Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, J., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. and Hurles, M.E. (2006) Global variation in copy number in the human genome. *Nature* **444**, 444-454.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199-204.

Renieri, A., Meloni, I., Longo, I., Ariani, F., Mari, F., Pescucci, C. and Cambi, F. (2003) Rett syndrome: the complex nature of a monogenic disease. *J Mol Med* **81**, 346-354.

Riazi, A.M., Lee, H., Hsu, C. and Van Arsdell, G. (2005) CSX/Nkx2.5 modulates differentiation of skeletal myoblasts and promotes differentiation into neuronal cells in vitro. *J Biol Chem* **280**, 10716-10720.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L. and et, a.l. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066-1073.

Risch, A. and Plass, C. (2008) Lung cancer epigenetics and genetics. *Int J Cancer* **123,** 1-7.

Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffith, O.L., Zhang, X., Pan, Y., Hassel, M., Sleumer, M.C., Pan, W., Pleasance, E.D., Chuang, M., Hao, H., Li, Y.Y., Robertson, N., Fjell, C., Li, B., Montgomery, S.B., Astakhova, T., Zhou, J., Sander, J., Siddiqui, A.S. and Jones, S.J. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* **34**, D68-73.

Rodd, Z.A., Bertsch, B.A., Strother, W.N., Le-Niculescu, H., Balaraman, Y., Hayden, E., Jerome, R.E., Lumeng, L., Nurnberger, J.I. Jr, Edenberg, H.J., McBride, W.J. and Niculescu, A.B. (2007) Candidate genes, pathways and mechanisms for alcoholism: an expanded convergent functional genomics approach. *Pharmacogenomics J* **7**, 222-256.

Ropers, H.H. (2006) X-linked mental retardation: many genes for a complex disorder. *Curr Opin Genet Dev* **16**, 260-269.

Ropers, H.H. and Hamel, B.C. (2005) X-linked mental retardation. *Nat Rev Genet* **6**, 46-57.

Rossi, S., Masotti, D., Nardini, C., Bonora, E., Romeo, G., Macii, E., Benini, L. and Volinia, S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes. *Nuc Acid Res* **34**, W285-W292

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386.

Ruta. M. , Howk, R., Ricca, G., Drohan, W., Zabelshansky, M., Laureys, G., Barton, D.E., Francke, U., Schlessinger, J. and Givol, D. (1988) A novel protein tyrosine kinase gene whose expression is modulated during endothelial cell differentiation. *Oncogene* **3**, 9-15.

Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y. and Lancet, D. (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res* **31**, 142-146.

Saito, T., Kinoshita, A., Yoshiura, K.i., Makita, Y., Wakui, K., Honke, K., Niikawa, N. and Taniguchi, N. (2001) Domain-specific mutations of a transforming growth factor (TGF)-beta 1 latency-associated peptide cause Camurati-Engelmann disease because of the formation of a constitutively active form of TGF-beta 1. *J Biol Chem* **276**, 11469-11472.

Salonen, J.T., Uimari, P., Aalto, J.M., Pirskanen, M., Kaikkonen, J., Todorova, B., Hypponen, J., Korhonen, V.P., Asikainen, J., Devine, C., Tuomainen, T.P., Luedemann, J., Nauck, M., Kerner, W., Stephens, R.H., New, J.P., Ollier, W.E., Gibson, J.M., Payton, A., Horan, M.A., Pendleton, N., Mahoney, W., Meyre, D., Delplanque, J., Froguel, P., Luzzatto, O., Yakir, B. and Darvasi, A. (2007) Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. *Am J Hum Genet* **81**, 338-345.

Sambrook, J. and Russell, D.W. (2001) *Molecular cloning: A laboratory manual*, Third Edition Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Sampson, P.D., Streissguth, A.P., Bookstein, F.L., Little, R.E., Clarren, S.K., Dehaene, P., Hanson, J.W. and Graham, J.M. (1997) Incidence of fetal alcohol syndrome and prevalence of alcohol-related neurodevelopmental disorder. *Teratology* 317-326.

Schlief, M.L. and Gitlin, J.D. (2006) Copper homeostasis in the CNS: a novel link between the NMDA receptor and copper homeostasis in the hippocampus. *Mol Neurobiol* **33**, 81-90.

Scuteri, A., Sanna, S., Chen, W.M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orru, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G.B., Fink, A.A., Weder, A.B., Cooper, R.S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E. and Abecasis, G.R. (2007) Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genet* **3**, e115.

Shah, M., Foreman, D.M. and Ferguson, M.W. (1995) Neutralisation of TGF-beta 1 and TGF-beta 2 or exogenous addition of TGF-beta 3 to cutaneous rat wounds reduces scarring. *J Cell Sci* **108 (Pt 3)**, 985-1002.

Shahbazian, M.D. and Zoghbi, H.Y. (2002) Rett syndrome and MeCP2: linking epigenetics and neuronal function. *Am J Hum Genet* **71,** 1259-1272.

Shashi, V., Berry, M.N., Shoaf, S., Sciote, J.J., Goldstein, D. and Hart, T.C. (2000) A unique form of mental retardation with a distinctive phenotype maps to Xq26-q27. *Am J Hum Genet* **66**, 469-479.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311.

Shi, Y. and Massague, J. (2003) Mechanisms of TGF-beta signalling from cell membrane to the nucleus. *Cell* **113**, 685-700.

Shiraishi, M. and Hayatsu, H. (2004) High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. *DNA Res* **11,** 409-415.

Shrimpton, A.E., Braddock, B.R. and Hoo, J.J. (2000) Narrowing the map of a gene (MRXS9) for X-linked mental retardation, microcephaly, and variably short stature at Xq12-q21.31. *Am J Med Genet* **92**, 155-156.

Shrimpton, A.E., Daly, K.M. and Hoo, J.J. (1999) Mapping of a gene (MRXS9) for X-linked mental retardation, microcephaly, and variably short stature to Xq12-q21.31. *Am J Med Genet* **84**, 293-299.

Smith, N.G. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions. *Gene* **318**, 169-175.

Sokol, R.J., Ager, J., Martier, S., Debanne, S., Ernhart, C., Kuzma, J. and Miller, S.I. (1986) Significant determinants of susceptibility to alcohol teratogenicity. *Ann N Y Acad Sci* **477**, 87-102.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1994) The transmission/disequilibrium test detects cosegregation and linkage. *Am J Hum Genet* **54**, 559-560.

Steinmuller, R., Steinberger, D. and Muller, U. (1998) MEHMO (mental retardation, epileptic seizures, hypogonadism and -genitalism, microcephaly, obesity), a novel syndrome: assignment of disease locus to xp21.1-p22.13. *Eur J Hum Genet* **6**, 201-206.

Stephenson, S.E., Dubach, D., Lim, C.M., Mercer, J.F. and La Fontaine, S. (2005) A single PDZ domain protein interacts with the Menkes copper ATPase, ATP7A. A new protein implicated in copper

homeostasis. *J Biol Chem* **280**, 33270-33279.

Stevenson, R.E., Hane, B., Arena, J.F., May, M., Lawrence, L., Lubs, H.A. and Schwartz, C.E. (1997) Arch fingerprints, hypotonia, and areflexia associated with X linked mental retardation. *J Med Genet* **34**, 465-469.

Stockard, C.R. (1910) The influence of alcohol and other anaesthetics onembryonic development. *Am J Anat* **10**, 369-392.

Stoler, J.M., Ryan, L.M. and Holmes, L.B. (2002) Alcohol dehydrogenase 2 genotypes, maternal alcohol use, and infant outcome. *J Pediatr* **141**, 780-785.

Stratton, K.R., Howe, C.J. and Battaglia, F.C. (1996) *Fetal Alcohol Syndrome: Diagnosis, Epidemiology, Prevention, and Treatment*, Washington, DC: National Academy Press.

Strauch, K., Fimmers, R., Baur, M.P. and Wienker, T.F. (2003) How to model a complex trait. 1. General considerations and suggestions. *Hum Hered* **55**, 202-210.

Streissguth, A.P. and Dehaene, P. (1993) Fetal alcohol syndrome in twins of alcoholic mothers: concordance of diagnosis and IQ. *Am J Med Genet* **47**, 857-861.

Sulik, K.K. (2005) Genesis of alcohol-induced craniofacial dysmorphism. *Exp Biol Med (Maywood)* **230**, 366-375.

Sulik, K.K. and Johnston, M.C. (1983) Sequence of developmental alterations following acute ethanol exposure in mice: craniofacial features of the fetal alcohol syndrome. *Am J Anat* **166**, 257-269.

Sulik, K.K., Johnston, M.C. and Webb, M.A. (1981) Fetal alcohol syndrome: embryogenesis in a mouse model. *Science* **214**, 936-938.

Sun, X. and Guo, B. (2006) Genotyping single-nucleotide polymorphisms by matrix-assisted laser desorption/ionization time-of-flight-based mini-sequencing. *Methods Mol Med* **128**, 225-230.

Surani, M.A. (1998) Imprinting and the initiation of gene silencing in the germ line. *Cell* **93**, 309-312.

Tabor, H.K., Risch, N.J. and Myers, R.M. (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **3**, 391-397.

Tassabehji, M., Hammond, P., Karmiloff-Smith, A., Thompson, P., Thorgeirsson, S.S., Durkin, M.E., Popescu, N.C., Hutton, T., Metcalfe, K., Rucka, A., Stewart, H., Read, A.P., Maconochie, M. and Donnai, D. (2005) GTF2IRD1 in craniofacial development of humans and mice. *Science* **310**, 1184-1187.

Tayaramma, T., Ma, B., Rohde, M. and Mayer, H. (2006) Chromatin-remodeling factors allow differentiation of bone marrow cells into insulin-producing cells. *Stem Cells* **24**, 2858-2867.

Tesseur, I., Zou, K., Esposito, L., Bard, F., Berber, E., Can, J.V., Lin, A.H., Crews, L., Tremblay, P., Mathews, P., Mucke, L., Masliah, E. and Wyss-Coray, T. (2006) Deficiency in neuronal TGF-beta signalling promotes neurodegeneration and Alzheimer's pathology. *J Clin Invest* **116**, 3060-3069.

The Wellcome trust case control consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678.

Thiery, J.P. (2003) Cell adhesion in development: a complex signalling network. *Curr Opin Genet Dev* **13**, 365-371.

Thomas, J.D., Burchette, T.L., Dominguez, H.D. and Riley, E.P. (2000) Neonatal alcohol exposure produces more severe motor coordination deficits in high alcohol sensitive rats compared to low alcohol sensitive rats. *Alcohol* **20**, 93-99.

Thomas, J.D., Melcer, T., Weinert, S. and Riley, E.P. (1998) Neonatal alcohol exposure produces hyperactivity in high-alcohol-sensitive but not in low-alcohol-sensitive rats. *Alcohol* **16**, 237-242.

Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M.A., Adeyemo, A., Patti, M.E., Semple, C.A. and Hide, W. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res* **34**, 3067-3081.

Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B. and Hide, W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* **33**, 1544-1552.

Tindall, E.A., Speight, G., Petersen, D.C., Padilla, E.J. and Hayes, V.M. (2007) Novel Plexor SNP genotyping technology: comparisons with TaqMan and homogenous MassEXTEND MALDI-TOF mass spectrometry. *Hum Mutat* **28**, 922-927.

Todd, J.A. (2006) Statistical false positive or true disease pathway? *Nat Genet* **38**, 731-733.

Tong, Z., Yang, Z., Meyer, J.J., McInnes, A.W., Xue, L., Azimi, A.M., Baird, J., Zhao, Y., Pearson, E., Wang, C., Chen, Y. and Zhang, K. (2006) A novel locus for X-linked retinitis pigmentosa. *Ann Acad Med Singapore* **35**, 476-478.

Tost, J. and Gut, I.G. (2007) DNA methylation analysis by pyrosequencing. *Nat Protoc* **2**, 2265-2275.

Tost, J. and Gut, I.G. (2005) Genotyping single nucleotide polymorphisms by MALDI mass spectrometry in clinical applications. *Clin Biochem* **38**, 335-350.

Toyota, M., Ho, C., Ahuja, N., Jair, K.W., Li, Q., Ohe-Toyota, M., Baylin, S.B. and Issa, J.P. (1999) Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res* **59,** 2307-2312.

Trump, D., Dixon, P.H., Mumm, S., Wooding, C., Davies, K.E., Schlessinger, D., Whyte, M.P. and Thakker, R.V. (1998) Localisation of X linked recessive idiopathic hypoparathyroidism to a 1.5 Mb region on Xq26-q27. *J Med Genet* **35**, 905-909.

Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T. and Sun, F. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7**, 31.

Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* **4**, R75.

Turner, G., Gedeon, A. and Mulley, J. (1994) X-linked mental retardation with heterozygous expression and macrocephaly: pericentromeric gene localization. *Am J Med Genet* **51**, 575-580.

Tzeng, J.Y., Devlin, B., Wasserman, L. and Roeder, K. (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* **72**, 891-902.

Valles, S., Pitarch, J., Renau-Piqueras, J. and Guerri, C. (1997) Ethanol exposure affects glial fibrillary acidic protein gene expression and transcription during rat brain development. *J Neurochem* **69**, 2484-2493.

van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A. and Brunner, H.G. (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet* **11**, 57-63.

Van Esch, H., Zanni, G., Holvoet, M., Borghgraef, M., Chelly, J., Fryns, J.P. and Devriendt, K. (2005) X-linked mental retardation, short stature, microcephaly and hypogonadism maps to Xp22.1-p21.3 in a Belgian family. *Eur J Med Genet* **48**, 145-52.

van Genderen, C., Okamura, R.M., Farinas, I., Quo, R.G., Parslow, T.G., Bruhn, L. and Grosschedl, R. (1994) Development of several organs that require inductive epithelial-mesenchymal interactions is impaired in LEF-1-deficient mice. *Genes Dev* **8**, 2691-2703.

Viljoen, D.L, Croxford, J., Gossage, J.P., Kodituwakku, P.W. and May, P.A. (2002) Characteristics of mothers of children with fetal alcohol syndrome in the Western Cape Province of South Africa: a case control study. *J Stud Alcohol* **63**, 6-17.

Viljoen, D.L., Carr, L.G., Foroud, T.M., Brooke, L., Ramsay, M. and Li, T.K. (2001) Alcohol dehydrogenase-2*2 allele is associated with decreased prevalence of fetal alcohol syndrome in the mixed-ancestry population of the Western Cape Province, South Africa. *Alcohol Clin Exp Res* **25**, 1719-1722.

Viljoen, D.L., Gossage, J.P., Brooke, L., Adnams, C.M., Jones, K.L., Robinson, L.K., Hoyme, H.E., Snell, C., Khaole, N.C., Kodituwakku, P., Asante, K.O., Findlay, R., Quinton, B., Marais, A.S., Kalberg, W.O. and May, P.A. (2005) Fetal alcohol syndrome epidemiology in a South African community: a second study of a very high prevalence area. *J Stud Alcohol* **66**, 593-604.

Vitale, E., Specchia, C., Devoto, M., Angius, A., Rong, S., Rocchi, M., Schwalb, M., Demelas, L., Paglietti, D., Manca, S., Mastropaolo, C. and Serra, G. (2001) Novel X-linked mental retardation syndrome with short stature maps to Xq24. *Am J Med Genet* **103**, 1-8.

Von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. and Bork, P. (2007) STRING 7: Recent developments in the integration and prediction of protein interactions. *Nuc Acid Res* **35**, D358-362.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33**, D433-437.

Wagschal, A. and Feil, R. (2006) Genomic imprinting in the placenta. *Cytogenet Genome Res* **113**, 90-98.

Walsh, F.S. and Doherty, P. (1997) Neural cell adhesion molecules of the immunoglobulin superfamily: role in axon growth and guidance. *Annu Rev Cell Dev Biol* **13**, 425-456.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. and Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831-845.

Warren, K.R. and Foudin, L.L. (2001) Alcohol-related birth defects--the past, present, and future. *Alcohol Res Health* **25**, 153-158.

Watase, K. and Zoghbi, H.Y. (2003) Modelling brain diseases in mice: the challenges of design and analysis. *Nat Rev Genet* **4**, 296-307.

Watty, A., Prieto, F., Beneyto, M., Neugebauer, M. and Gal, A. (1991) Gene localization in a family with X-linked syndromal mental retardation (Prieto syndrome). *Am J Med Genet* **38**, 234-239.

Weir, B.S., Anderson, A.D. and Hepler, A.B. (2006) Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**, 771-780.

Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat Genet* **26**, 151-157.

Welch-Carre, E. (2005) The neurodevelopmental consequences of prenatal alcohol exposure. *Adv Neonatal Care* **5**, 217-229.

West, J.R., Goodlett, C.R. and Brandt, J.P. (1990) New approaches to research on the long-term consequences of prenatal exposure to alcohol. *Alcohol Clin Exp Res* **14**, 684-689.

Wieland, I., Jakubiczka, S., Muschke, P., Cohen, M., Thiele, H., Gerlach, K.L., Adams, R.H. and Wieacker, P. (2004) Mutations of the ephrin-B1 gene cause craniofrontonasal syndrome. *Am J Hum Genet* **74**, 1209-1215.

Wijker, M., Ligtenberg, M.J., Schoute, F., Defesche, J.C., Pals, G., Bolhuis, P.A., Ropers, H.H., Hulsebos, T.J., Menko, F.H., van Oost, B.A. and et, a.l. (1995) The gene for hereditary bullous dystrophy, X-linked macular type, maps to the Xq27.3-qter region. *Am J Hum Genet* **56**, 1096-1100.

Wilkemeyer, M.F., Chen, S.Y., Menkari, C.E., Brenneman, D.E., Sulik, K.K. and Charness, M.E. (2003) Differential effects of ethanol antagonism and neuroprotection in peptide fragment NAPVSIPQ prevention of ethanol-induced developmental toxicity. *Proc Natl Acad Sci U S A* **100**, 8543-8548.

Wilson, M., Mulley, J., Gedeon, A., Robinson, H. and Turner, G. (1991) New X-linked syndrome of mental retardation, gynecomastia, and obesity is linked to DXS255. *Am J Med Genet* **40**, 406-413.

Wittwer, B., Kircheisen, R., Leutelt, J., Orth, U. and Gal, A. (1996) New X-linked mental retardation syndrome with the gene mapped tentatively in Xp22.3. *Am J Med Genet* **64**, 42-49.

Xu, J., Yeon, J.E., Chang, H., Tison, G., Chen, G.J., Wands, J. and de la Monte, S. (2003) Ethanol impairs insulin-stimulated neuronal survival in the developing brain: role of PTEN phosphatase. *J Biol Chem* **278**, 26929-26937.

Yang, T., Pfister, M., Blin, N., Zenner, H.P., Pusch, C.M. and Smith, R.J. (2005) Genetic heterogeneity of deafness phenotypes linked to DFNA4. *Am J Med Genet A* **139**, 9-12.

Yuan, H.Y., Chiou, J.J., Tseng, W.H., Liu, C.H., Liu, C.K., Lin, Y.J., Wang, H.H., Yao, A., Chen, Y.T. and Hsu, C.N. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res* **34**, W635-641.

Yue, P. and Moult, J. (2006) Identification and analysis of deleterious human SNPs. *J Mol Biol* **356**, 1263-1274.

Yue, P., Li, Z. and Moult, J. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* **353**, 459-473.

Yue, P., Melamud, E. and Moult, J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**, 166.

Zhu, D., Alcorn, D.M., Antonarakis, S.E., Levin, L.S., Huang, P.C., Mitchell, T.N., Warren, A.C. and Maumenee, I.H. (1990) Assignment of the Nance-Horan syndrome to the distal short arm of the X chromosome. *Hum Genet* **86**, 54-58.

Internet resources

The following online resources were used for analysis in this body of work:

dbSNP	http:// www.ncbi.nlm.nih.gov/SNP (Accessed July 2007)
Dragon Disease Explorer	http://research.i2r.a-star.edu.sg/DRAGON/DE/ (Accessed April 2006, July 2007)
DTFAM	http://research.i2r.a-star.edu.sg/DRAGON/ TFAM_v2/index.html (Accessed April 2006, July 2007)
Ensembl	http://www.ensembl.org (Accessed September 2005 – December 2007)
FastSNP	http://fastsnp.ibms.sinica.edu.tw/ (Accessed July 2007)
MGD	http://www.informatics.jax.org (Accessed April 2006, July 2007)
Primer3	http://frodo.wi.mit.edu/ (Accessed November 2006)
TFSearch	http://www.rwcp.or.jp/papia/ (Accessed April 2007)
UCSC genome browser	http://genome.ucsc.edu/ (Accessed November 2006 – May 2007)
UCSC In-silico PCR	http://genome.ucsc.edu/cgi-bin/hgPcr?command=start (Accessed November 2006)
Uniprot database	http://www.ebi.uniprot.org/ (Accessed December 2007)
US NLM - PubMed tutorial	http://www.nlm.nih.gov/bsd/disted/pubmed.html

Addendum A

Python Scripts

A.1 Generating the binary matrix

```
def file_to_list(file):
	list = []
	for term in open(file):
		term = term.strip()
		list.append(term)
	return list

import datetime
start_time = datetime.datetime.now()

file1 = "/home/zane/RESULTS/Text-mining_M2/Binary_Start.txt"
list1 = file_to_list(file1)
D = {}

for x in list1:
	D[x] = 0

file2 = "/home/zane/RESULTS/Binary_Clustering/CATEGORIES/microcephaly.txt"
list2 = file_to_list(file2)
change = 1
count = 0

for id in list2:
	if id in D.keys():
		D[id] = change
		count = count + 1
Dlist = D.keys()
Dlist.sort()

print "terms present in both lists:", count
print "The dictionary in alphabetical order"
for x in Dlist:
	print x, "\t", D[x]

end_time = datetime.datetime.now()
print "The time taken for this process is:", (end_time - start_time),"\n

"
```

A.2 Scripts adapted from Tiffin et al. (2005) used to generate the initial candidate gene list *(Alterations are shown in Italics).*

```
import datetime
start_time = datetime.datetime.now()

import real_disease_gene_finder
import create_okc

disease_names = ["fetal_alcohol_syndrome"]
all_diseases_results_list = [['total_abstracts','mismatch',
'no_of_final_terms', 'common_hugo_list', 'unknown_transcripts']]
total_no_iterations = 0
x = 1
y = 2

for disease_name in disease_names:
 final_results_list = [['total_abstracts', 'mismatch',
'no_of_final_terms','common_hugo_list', 'unknown_transcripts']]
```

```
 no_of_iterations = 0

 for mismatch in range(x,y):
   m = 5 #mismatch + 1
   n = 6
   for no_of_final_terms in range (m,n):

       results_list =
       real_disease_gene_finder.real_disease_gene_finder(disease_name,
       no_of_final_terms, mismatch)
   no_of_iterations = no_of_iterations + 1
   total_no_iterations = total_no_iterations + 1
   final_results_list.append(results_list)
   all_diseases_results_list.append(results_list)

 create_okc.create_okc(disease_name, no_of_iterations, final_results_list)

end_time = datetime.datetime.now()
print "The time taken for this process is:", (end_time - start_time),"\n"
```

real disease gene finder module

```
def real_disease_gene_finder(disease_name, no_of_final_terms, mismatch):
 print 'disease is', disease_name

 import datetime
 start_time = datetime.datetime.now()

 import MySQLdb as dbi

 terms = "anatomy_terms"
 annot_frequency = "refseq_gene_count_frequency"
 delimiter = "\t"
 filename = file("hugo_id.txt", "a")
 filename2 = file("unknown_transcript_id.txt", "a")
 filename3 = file("term_count.txt", "a")
 print>>filename, disease_name
 print>>filename2, disease_name
 print>>filename3, disease_name

 assoc_terms_list = []
 dbc = dbi.connect(host="localhost", db="disease_terms_assoc", user="zane")
 cursor = dbc.cursor()
 cursor.execute("""SELECT DISTINCT trim(anatomy_terms),frequency_assoc FROM
%s;""" % disease_name)

 row = cursor.fetchone()
 while row is not None:
       print row
  assoc_term = row[0]
  assoc_term = assoc_term.strip()
  assoc_term = assoc_term.lower()
  if assoc_term != '':
   assoc_terms_list.append(assoc_term)
  row = cursor.fetchone()
 cursor.close()
 dbc.close()

 print 'associated terms are:\n', assoc_terms_list
 print 'no. of terms is', len(assoc_terms_list)
```

```
 assoc_terms_freq_list = []
 assoc_single_terms_freq =[]

 dbc = dbi.connect(host="localhost", db="term_annotation_frequency",
user="zane")

 for assoc_term in assoc_terms_list:

  cursor = dbc.cursor()
  cursor.execute("""SELECT DISTINCT trim(term) FROM all_terms_frequency
      where trim(term) like "%%%s%%";""" % (assoc_term))

  row = cursor.fetchone()

  while row is not None:
   assoc_freq_term = row[0]
   if "|" in assoc_freq_term:
    #print 'freq_term is', assoc_freq_term
    assoc_freq_term = assoc_freq_term.lower()
    synonyms = assoc_freq_term.split("|")

    if assoc_term in synonyms:
     assoc_terms_freq_list.append(assoc_freq_term)

     for term in synonyms:
      term = term.strip()
      term = term.lower()
      #print 'term is', term
      if term not in assoc_single_terms_freq:
       assoc_single_terms_freq.append(term)
      else:
       print 'The term', term, 'is already there'

      print 'assoc_single_terms_freq is', assoc_single_terms_freq
    else:
     print 'This term is not valid:', assoc_freq_term
   else:
    print 'assoc_freq_term unsplit is', assoc_freq_term
    assoc_freq_term = assoc_freq_term.strip()
    assoc_freq_term = assoc_freq_term.lower()

    if assoc_term == assoc_freq_term and assoc_freq_term not in
assoc_single_terms_freq:
     assoc_single_terms_freq.append(assoc_freq_term)
     print 'this simple term is present', assoc_term
    else:
     print 'this simple term is excluded (already there or not equivalent)',
assoc_term

    if assoc_term == assoc_freq_term and assoc_freq_term not in
assoc_terms_freq_list:
     assoc_terms_freq_list.append(assoc_freq_term)
   row = cursor.fetchone()
  cursor.close()
 dbc.close()

 print 'Compound terms equal to assoc terms are:\n', assoc_terms_freq_list
 print 'Single terms equal to assoc terms are:\n',assoc_single_terms_freq
```

```
 dbc = dbi.connect(host="localhost", db="disease_terms_assoc", user="zane")
 freq_assoc_dict = {}

 for term in assoc_terms_freq_list:
  if "|" in term:
   print 'compound term is', term
   freq_assoc_dict[term] = 0
   print 'frequency of term is', freq_assoc_dict[term]
   single_terms_list = term.split("|")
   print 'single terms list is', single_terms_list
   for single_term in single_terms_list:
    if single_term in assoc_terms_list:
     print 'single term in assoc_terms_list is', single_term
     cursor = dbc.cursor()
     cursor.execute("""SELECT DISTINCT trim(anatomy_terms), frequency_assoc
          FROM %s
          where anatomy_terms = "%s"; """ % (disease_name, single_term))

     row = cursor.fetchone()
     while row is not None:
      print 'records returned are', row[0], ',', row[1]

      freq_assoc_dict[term] = freq_assoc_dict[term] + row[1]
      print 'new frequency of association value for', term, 'is',
freq_assoc_dict[term]
 else:
 print 'the term', term,'is already in the dictionary'
      row = cursor.fetchone()
     cursor.close()
    else:
     print 'The term', single_term, 'in not an associated term'
  else:
   print 'there is no pipe in', term

   cursor2 = dbc.cursor()
   cursor2.execute("""SELECT DISTINCT trim(anatomy_terms), frequency_assoc
       FROM %s
       where anatomy_terms = "%s"; """ % (disease_name, term))
   row = cursor2.fetchone()
   while row is not None:
    #print row[0], row[1]
    freq_assoc_dict[row[0]] = row[1]
    row = cursor2.fetchone()

   cursor2.close()
 dbc.close()

 print 'frequency of association dictionary is', freq_assoc_dict
 assoc_keys_list = freq_assoc_dict.keys()
 print "the number of keys in freq_assoc_dict is", len(assoc_keys_list)

 freq_annot_dict = {}

 dbc = dbi.connect(host="localhost", db="term_annotation_frequency",
user="zane")

 for term in assoc_keys_list:
  #print "key term is", term
  cursor = dbc.cursor()
  cursor.execute("""SELECT DISTINCT trim(term), refseq_gene_count_frequency
```

```
      FROM all_terms_frequency
      WHERE trim(term) = "%s";""" % (term))

 row = cursor.fetchone()
 while row is not None:
  print 'retrieved data is', row[0], row[1]
  annot_term = row[0]
  annot_term = annot_term.lower()
  annot_term = annot_term.strip()
  freq_annot_dict[annot_term]= row[1]
  row = cursor.fetchone()
 cursor.close()
dbc.close()

print 'frequency of annotation dictionary is', freq_annot_dict
annot_keys_list = freq_annot_dict.keys()
print 'length of annot dict is', len(annot_keys_list)

common_keys_list = []
for key in annot_keys_list:
 if key in assoc_keys_list:
  common_keys_list.append(key)
 else:
  print 'this key', key,'is not in common'
print 'common keys are', common_keys_list
for term in assoc_terms_freq_l

ranked_term_dict = {}

for key in common_keys_list:

 assoc_freq = freq_assoc_dict[key]
 annot_freq = freq_annot_dict[key]
 score = ((2*assoc_freq) + annot_freq)/2

 print 'score for key', key,'is', score

 ranked_term_dict[key] = score
 ranked_keys_list = ranked_term_dict.keys()

print 'ranked term dict is', ranked_term_dict
print 'there are',len(ranked_keys_list), 'ranked terms'

final_term_list = []

tissue_tuples = [(v,k) for k,v in ranked_term_dict.items()]
tissue_tuples.sort()
tissue_tuples.reverse()

for value in tissue_tuples:
 final_term_list.append(value[1])
print "the final term list is:", final_term_list
print "the final term list contains", len(final_term_list), "terms"
short_term_list = final_term_list[:no_of_final_terms]
print "the short list is:", short_term_list

dbc = dbi.connect(host= "localhost", db= "ensembl_mart_33", user = "zane")
final_term_list
short_term_list
```

```
 id_list_dict = {}

 for term in short_term_list:
  id_list = []

  if "|" in term:
   print "The term contains synonyms"

   print "complex term is", term
   terms = term.split("|")
   for single_term in terms:

    single_term = single_term.lower()
    single_term = single_term.strip()
    #print 'single_term is', single_term

    cursor = dbc.cursor()
    cursor.execute("""SELECT DISTINCT transcript_id_key,
anatomical_system_description
      FROM hsapiens_gene_ensembl__expression_est_descriptions__dm
      WHERE anatomical_system_description like "%%%s%%";""" % (single_term))
    row = cursor.fetchone()
    while row is not None:
     new_anat_terms1 = []

     print row[0],row[1]
     anat_terms = row[1]
     single_anat_terms = anat_terms.split("->")

     for anat_term1 in single_anat_terms:
      anat_term1 = anat_term1.strip()
      anat_term1 = anat_term1.lower()
      if anat_term1 not in new_anat_terms1:
       new_anat_terms1.append(anat_term1)

     print "single_anat_terms in list are", new_anat_terms1

     if single_term in new_anat_terms1:
      print 'the id', row[0],'is valid'
      id_list.append(row[0])
     else:
      print 'this id is not valid'
     row = cursor.fetchone()
    cursor.close()

  else:
   print "The term", term, "contains no synonyms"
   cursor2 = dbc.cursor()

   cursor2.execute("""SELECT DISTINCT transcript_id_key,
anatomical_system_description
        FROM hsapiens_gene_ensembl__expression_est_descriptions__dm
        WHERE anatomical_system_description like "%%%s%%";""" % (term))
   row = cursor2.fetchone()
   while row is not None:
    new_anat_terms2 = []
    print row[0],row[1]
    anat_terms = row[1]
    print 'anat_terms are', anat_terms
    single_anat_terms = anat_terms.split("->")
```

```
    for anat_term in single_anat_terms:
     anat_term = anat_term.strip()
     anat_term = anat_term.lower()
     print 'anat_term is', anat_term
     if anat_term not in new_anat_terms2:
      new_anat_terms2.append(anat_term)

    if term in new_anat_terms2:
     print 'the id is valid'
     id_list.append(row[0])
    else:
     print 'the id is not valid'

    row = cursor2.fetchone()
   cursor2.close()
  print 'id_list is', id_list
  id_list_dict[term] = id_list

 dbc.close()
 print 'id_list dictionary is', id_list_dict

no_of_final_terms
 mismatch
 all_id_list = []

 for key in id_list_dict:
  id_list = id_list_dict[key]

  for id in id_list:
   if id not in all_id_list:
    all_id_list.append(id)
  print 'for', key, 'the length of the id_list is', len(id_list)
 print 'all_id_list is', all_id_list, 'and it contains',
len(all_id_list),'ids'

 common_id_list = []
 for all_id in all_id_list:
  n=0
  for key in id_list_dict:
   id_list = id_list_dict[key]
   if all_id in id_list:
    print 'all_id present in id_list with key', key
    n=n+1
  print all_id,'has a count of',n
  if n >= no_of_final_terms - mismatch:
   print 'append id'
   common_id_list.append(all_id)
   print >> filename3, all_id, n
 print 'common_id_list is', len(common_id_list), 'and contains',
common_id_list

 common_hugo_list = []
 unknown_transcripts = []

 for id in common_id_list:
  dbc = dbi.connect(host = "localhost", db = "ensembl_mart_33", user =
"zane")
  cursor = dbc.cursor()
  cursor.execute("""select distinct display_id_list, transcript_id_key
       from hsapiens_gene_ensembl__xref_hugo__dm
```

```
       where transcript_id_key = %s;""" % (id))

  row = cursor.fetchone()
  while row is not None:
   print 'display id in row[0] is', row[0]
   if row[0] != '' and row[0] not in common_hugo_list:
    common_hugo_list.append(row[0])
    print >> filename, row[0]
   if row[0] == '' and row[1] not in unknown_transcripts:
    unknown_transcripts.append(row[1])
    print >> filename2, row[1]
   row = cursor.fetchone()
  cursor.close()
  dbc.close()

 common_hugo_list.sort()

 print 'common hugo list contains', len(common_hugo_list),'genes, and is',
common_hugo_list
 print 'there are', len(unknown_transcripts), 'unknown transcripts, which
are', unknown_transcripts

  dbc = dbi.connect(host="localhost", db="disease_terms_assoc", user="zane")

 cursor = dbc.cursor()
 cursor.execute("""SELECT DISTINCT abstracts_disease FROM %s
 WHERE abstracts_disease>0;"""
 % (disease_name))

 row = cursor.fetchone()
 while row is not None:
  total_abstracts = row[0]
  row = cursor.fetchone()
 cursor.close()
 dbc.close()

 results_list = [total_abstracts, mismatch, no_of_final_terms,
len(common_hugo_list), len(unknown_transcripts)]
 #print results_list

 end_time = datetime.datetime.now()
 #print "The time taken for this process is:", (end_time - start_time),"\n"

 return results_list
```

Create okc module

```
def create_okc(disease_name, no_of_iterations, final_results_list):
 import string
 file_title = "/home/zane/RESULTS/Gene_extract_results/" + str(disease_name)
+ '.okc'

 filename = file(file_title, "a")

 field_names = final_results_list[0]
 print field_names
 print "filename is", file_title

 print >> filename, "NEW RUN"
 print >> filename, len(field_names)," ",no_of_iterations
```

```
 for field_name in field_names:
  print field_name
  print >> filename, field_name

 max_list = [0 for field_number in range(len(final_results_list[1]))]
 min_list = [100000 for field_number in range(len(final_results_list[1]))]

 for results in final_results_list[1:]:
  for field_number in range(len(results)):
   if results[field_number] > max_list[field_number]:
    max_list[field_number] = results[field_number]
   if results[field_number] < min_list[field_number]:
    min_list[field_number] = results[field_number]
 print "max_list", max_list
 print "min_list", min_list

 for field_number in range(len(min_list)):
  print >> filename, min_list[field_number]," ",max_list[field_number]," ",3

 for results in final_results_list[1:]:
  for element in results:
   print >> filename, element," ",
  print >> filename
```

A.3 Scripts used to obtain human orthologues of mouse genes from a locally installed copy of the MGI database for the binary analysis

```
def file_to_list(file):
      list = []
      for term in open(file):
            term = term.strip()
            list.append(term)
      return list

def list_to_file(list):
      file =
open("/home/zane/RESULTS/Binary_Clustering/CATEGORIES/3B.Mouse_timin
g/pre-embryo.txt","w")
      for term in list:
            file.write(term)
            file.write("\n")
      file.close()

import datetime
start_time = datetime.datetime.now()
import MySQLdb as dbi

list = []
human_list = []
count_human = 0

file = "/home/zane/RESULTS/MGD_results/Development_stage/mouse/Non-
redundant/pre-embryo.txt"
list = file_to_list(file)

for term in list:
```

```
      dbc = dbi.connect(host="localhost",db="MGI__3_4",user="zane")
      cursor = dbc.cursor()
      cursor.execute("SELECT DISTINCT trim(human_symbol) FROM
human_mouse_orthology WHERE trim(mouse_symbol) = \""+term+"\"")
      row = cursor.fetchone()

      while row is not None:
            assoc_term = row[0]
            assoc_term = assoc_term.strip()
            if assoc_term not in human_list:
                  count_human = count_human + 1
                  human_list.append(assoc_term)
            row = cursor.fetchone()
      cursor.close()
      dbc.close()

list_to_file(human_list)
end_time = datetime.datetime.now()
print "There are ", count_human, " human genes."
print "The time taken for this process is:", (end_time -
start_time),"\n"
```

A.4 Creating a non-redundant gene list

```
def file_to_list(file):
      list = []
      for term in open(file):
            term = term.strip()
            list.append(term)
      return list

def list_to_file(list):
      file = open("/home/zane/RESULTS/GO_terms/NR_310106.txt","w")
      for term in list:
            file.write(term)
            file.write("\n")
      file.close()

import datetime
start_time = datetime.datetime.now()

list = []
new_list = []
term_count = 0
count = 0
false_count = 0

file = "/home/zane/RESULTS/GO_results/All_terms_310103.txt"

list = file_to_list(file)

for term in list:
      term_count = term_count + 1
```

```
      if term in new_list:
            false_count = false_count + 1
      else:
            new_list.append(term)
            count = count + 1

list_to_file(new_list)

end_time = datetime.datetime.now()
print "There are ", term_count, "GO terms in the list."
print "There are ", count, "GO terms in the non-redundant list."
print "There were ", false_count, "redundant terms in the list."
print "The time taken for this process is:", (end_time -
start_time),"\n"
```

A.5 Generating a randomly selected gene list

```
import random, sys

file = open(fileName)
geneList = []

for line in file:
      geneList.append(line)

randomList = []
randomChoice = ""
counter = 0
while counter < numOfRandGenes:
      randomChoice = random.choice(geneList)
      if randomChoice not in randomList:
            randomList.append(randomChoice)
            counter+=1

      else:
            print "Duplicate excluded"

outputFile = open("/home/zane/random.txt",'a')
for entry in randomList:
      outputFile.write(entry)
outputFile.close
```

Addendum B

Additional Computational Results

Table B.1: 87 top-ranked genes for FASD identified using binary matrix filtering**.**

Criteria Matched	HUGO	Locus	Function
17/29	*FGFR1*	8p11.2	Involved in limb induction, play a role in bone elongation modulation
16/29	*MSX1*	4p16.3-p16.1	Potential repressor function in cell cycle progression, transcription repressor
15/29	*FGFR2*	10q26	Involved in vertebral development, important regulator of bone formation and osteoblast activity
15/29	*FOXG1B*	14q13	Embryonic transcriptional regulator, playing a critical role in brain development
15/29	*HOXA1*	7p15.3	Involved in the placement of hindbrain segments in the proper location along the anterior-posterior axis during development
14/29	*BMP4*	14q22-q23	Regulating myogenesis through dosage-dependent PAX3 expression in pre-myogenic cells, inducing apoptosis and chondrogenesis in the chick limb bud
14/29	*FGFR3*	4p16.3	Negative regulator of bone growth promotion, inhibition of chondrocyte proliferation and differentiation depending on devlopmental time
14/29	*GNAS*	20q13.2-q13.3	Involved as modulators or transducers in various transmembrane signalling systems primarily mediating the differential effects of parathyroid hormone
14/29	*PAX6*	11p13	Key regulator of eye, pancreas, central nervous system development and regulator of glial precursors in the ventral neural tube
13/29	*CASP3*	4q35	Effector caspase in the Fas proapoptotic pathway, playing a crucial role during apoptosis
13/29	*CITED2*	6q23.3	Cytokine-inducible transcription factor with transformation activity controlling left-right patterning and heart development
13/29	*DLX5*	7q21.3	Playing a role in forebrain and craniofacial development and in production of gaba-ergic neurons
13/29	*EGFR*	7p12	Involved in the control of cell growth and differentiation in early craniofacial development and palate closure and in keratinocyte differentiation
13/29	*GLI2*	2q24	Involved in the formation of lung, trachea and oesophagus, playing a role in head development
13/29	*GLI3*	7p14.1-p13	Involved in the development of the CNS, craniofacial structure, lung, trachea and oesophagus
13/29	*IGF2*	11p15.5	Potent mitogen influenced by placental lactogen and may be playing a role in fetal development
13/29	*ITGA6*	2q22-q31	Cell surface adhesion receptor mediating cell-adhesion to extra cellular matrix or to other cells, also involved in the fertilization and embryonic development
13/29	*MECP2*	Xq28	Involved in the regulation of gene expression (in normal neuronal maturation) and may regulate the transcription genes in neuronal cells, important in synapse development and neuronal plasticity
13/29	*MEST*	7q32	Mesoderm-specific transcript, expressed in fetal tissues from the paternal chromosome
13/29	*PAFAH1B1*	17p13.3	Component of overlapping but distinct signal pathways including DCX, that promotes neuronal migration
13/29	*PAX3*	2q36.1-q36.2	Involved in neurogenesis, in skeletal muscle development and in melanogenesis through MITF transactivation and maybe other processes
13/29	*PTCH*	9q22.32-q22.33	Repressing SMO activity in the absence of SHH, and involved in endocytosis and vesicle transport
13/29	*PTPN11*	12q24.1	Involved in intracellular signal transduction in response to PDGF, EGF and insulin
13/29	*RB1*	13q14.2	Regulator of cell growth, may function in cell cycle exit, differentiation and survival of hair cells
13/29	*SOX2*	3q26.3-q27	Activating FGF4 and modulator of LINE retroposons promoter activity
13/29	*TGFB1*	19q13.2	Stimulating articular chondrocyte cell growth through MAPK3 activation and inducing apoptosis in endothelial cells
12/29	*CREBBP*	16p13.3	Playing a pivotal role in embryonic development, involved in a variety of transcriptional pathways through chromatin remodeling
12/29	*CTNNB1*	3p22-p21.3	Multifunctional protein participating in cell-cell adhesion and Wnt-stimulated transcriptional activation and the establishment of a bipolar mitotic spindle

Criteria Matched	HUGO	Locus	Function
12/29	*DTNA*	18q12.1-q12.2	May be involved in signal transduction in myeloid cells during induction of granulocytic differentiation and/or at the commitment stage of differentiation or phagocytic cells
12/29	*GABRB3*	15q11.2-q12	Mediating neuronal inhibition by binding to the gaba-benzodiazepine receptor and opening an integral chloride channel
12/29	*JAG1*	20p12.1-p11.23	NOTCH1 ligand playing a pivotal role in the development of the organ of Corti and specification of some vestibular sensory epithelia
12/29	*MET*	7q31.2-q31.3	Important regulator of cell proliferation and differentiation, organ regeneration, embryogenesis and tumorogenesis
12/29	*NGFR*	17q21.31	Playing a central role for mediating inhibitory signals from CNS myelin
12/29	*PITX2*	4q25-q27	RNA polymerase 2 transcription factor, involved in the regulation of heart
12/29	*PXMP3*	8q21.1	Implicated in the biogenesis of peroxisomes
12/29	*SIX3*	2p21-p16	Involved in regulation of eye and neural plate development
12/29	*SMAD2*	18q21	Intracellular mediator of TGFB family of cytokines and activin type 1 receptor
12/29	*SNAP25*	20p12-p11.2	May play an important role in the synaptic function of specific neuronal systems
12/29	*UBE3A*	15q12	Playing a major role in neural development, role in regulating growth of neuronal processes or synapse formation
12/29	*VEGF*	6p12	Key regulator of blood vessel growth, and protecting endothelial cells from apoptosis
11/29	*AKT1*	14q32.32	A key regulator for cell growth, cell survival and metabolic insulin action
11/29	*APP*	21q21.2	Playing a significant role in regulating cerebral thrombosis and increases can profoundly enhance cerebral hemorrhage
11/29	*ARNT*	1q21	Dioxin receptor translocator, activating genes involved in metabolism, angiogenesis and apoptosis
11/29	*BCL2L1*	20q11.1	Involved in regulation of cell death by blocking the voltage dependent anion channel, # has a role in preventing BAX activation at the mitochondrial membrane
11/29	*BMP7*	20q13.2	Stimulating initial dendritic growth in sympathetic neurons, important regulator of cell development and differentiation of various organs
11/29	*BRCA1*	17q21	Negative regulator of mammary cell growth, having critical function in the proliferation and differentiation of neural progenitor cells
11/29	*CDKN1B*	12p13.1-p12	Critical terminal effectors of signal transduction pathways that control cell differentiation, mediating cell cycle regulation by AFX-like forkhead transcription factors
11/29	*COL2A1*	12q13.1	Involved in the structure of cartilage collagen
11/29	*CXCR4*	2q21	Allows cells to migrate in response to a gradient of chemokine ligands, playing a role of receptor for SDF1 to directing the primordial germ cell migration
11/29	*DCN*	12q13.2	Playing a role in epithelial/mesenchymal interactions during organ development and shaping, playing a role in matrix assembly
11/29	*EDG2*	9q31.3-q32	Involved in the control of myelination, activating phospholipase C
11/29	*EMX2*	10q26.1	Involved in brain development, normal growth and maturation of hippocampus
11/29	*FGF2*	4q26-q27	Inducer of anteroposterior neural pattern, essential for FGF8 and FGF10 reciprocal regulation in limb induction in mouse and involved in angiogenesis
11/29	*GDF11*	12q12	Involved in mesodermal formation and neurogenesis during embryonic development
11/29	*GJA1*	6q22.3	Member of the connexin of intercellular channels, providing a route for the diffusion of materials of low molecular weight from cell to cell
11/29	*GTF2I*	7q11.23	Multifunctional phosphoprotein with roles in transcription and signal transduction
11/29	*HMGB1*	13q12	DNA-binding protein, regulating gene transcription and stabilizing nucleosome formation
11/29	*HOXB2*	17q21.3	developmental regulatory system transcription factor that provide cells positional identities on the anterior-posterior axis
11/29	*LDB1*	10q24-q25	Required for LIM-homeodomain proteins to exert their biological activities

Criteria Matched	HUGO	Locus	Function
11/29	*LHX2*	9q33-q34.1	Acting as a tissue specific transcriptional activator of the alpha subunit of glycoprotein hormones, involved in the control of cell differentiation in developing lymphoid and neural cell types
11/29	*LIMK1*	7q11.23	May act as a link between stress-induced ceramide formation and reorganization of the actin cytoskeleton
11/29	*LMO4*	1p22.3	Transcriptional regulator participating to mammary gland development, repressor of BRCA1
11/29	*MAPK3*	16p11.2	Activation of MAPK3 is a key regulator of the increased transition to hypertrophic differentiation of the growth plate
11/29	*MARCKS*	6q21-q22.2	Actin filament cross linking protein regulator of actin cytoskeleton
11/29	*NCAM1*	11q23.1	Involved in neuron-neuron adhesion, neurite fasciculation, outgrowth of neurites
11/29	*NF1*	17q11.2	Required in endothelial cells but do not rule out a simultaneous requirement in the neural crest during cardiac development
11/29	*NPAS3*	14q12-q13	Neuronal transcription factor
11/29	*NR2F1*	5q14	Nuclear receptor involved in the organogenesis, transcription factor
11/29	*NR2F2*	15q26.1-q26.2	May be required for angiogenesis and heart development, negative post transcriptional regulator of MYOD1 function
11/29	*NR2F6*	19p13.1	Nuclear receptor involved in action regulation of the transcription of GNRH1 gene
11/29	*OTX1*	2p13	Required for sense organ development. Required for the refinement of exuberant axonal projections to subcortical targets
11/29	*PBX1*	1q23	May be playing a role in steroidogenesis, sexual development, and megakaryocytic gene expression
11/29	*PCGF4*	10p13	Involved in maintaining the transcriptional repressive state of genes, playing an essential role for the generation of self-renewing adult hematopoietic stem cells
11/29	*PDGFRA*	4q11-q12	Growth arrest specific gene (gas), subunit, receptor tyrosine kinase, class III, binding both PDGFA and PDGFB and having a tyrosine-protein kinase activity
11/29	*PHC1*	12p13	Can play a key role in organogenesis by helping to maintain the expression of a selector gene
11/29	*PHC2*	1p34.3	Required to maintain transcriptional repressive state of many genes, incl. HOX genes, during development
11/29	*PPARBP*	17q12-q21.1	Essential role for embryonic fibroblast differentiation pathway, and normal development of vital organ systems
11/29	*RARG*	12q13.13	Ligand activated transcription factor
11/29	*RXRA*	9q34.3	Involved in retinoic acid response pathway, regulates cardiac morphogenesis
11/29	*SOX11*	2p25.3	Playing a role in the developing nervous system
11/29	*SPRY2*	13q31-q32	Required for growth factor stimulated translocation of the protein to membrane ruffles
11/29	*TGFB2*	1q41	Having suppressing effects on interleukin-2 dependent T-cell growth
11/29	*TP53*	17p13.1	Transcriptional activator through acetylation of transactivation site by CREBBP, activator of target genes promoting growth arrest or cell death in response to DNA damage
11/29	*TWIST1*	7p21.2	Regulator of embryonic morphogenesis and playing an essential role in metastasis by promoting an epithelial-mesenchymal transition
11/29	*WNT5A*	3p21.1	Modulating cell fate and cell behavior during vertebrate development
11/29	*YWHAE*	17p13.3	Inhibitor of apoptosis through inhibiting the activation of p38 MAP kinase, multifunctional regulator required for cytoplasmic dynein function, neuronal migration and brain development
11/29	*ZIC2*	13q32	Putative regulator of the kinetic neural development

Table B.2: Known protein-protein interaction for the prioritized candidate genes obtained using STRING. The available evidence for the most significant interactions as well as the confidence score assigned for the interactions are shown.

Gene 1	Gene 2	CONFIDENCE SCORES					Combined confidence score
		Homology analysis	Coexpression evidence	Experimental data	Pathway data	Textmining co-occurrence	
TP53	*CREBBP*	0	0	0.991	0.800	0.758	0.999
FGFR3	*FGFB*	0	0	0.980	0.960	0.855	0.999
FGFR2	*FGFB*	0	0	0.997	0.960	0.869	0.999
FGFR1	*FGFB*	0	0	0.998	0.960	0.860	0.999
BCL2L1	*TP53*	0	0	0.985	0	0.911	0.998
PCGF4	*PHC2*	0	0	0.988	0	0.912	0.998
TGFB1	*DCN*	0	0	0.946	0.800	0.900	0.998
SMAD2	*TGFB1*	0	0	0	0.980	0.910	0.998
BRCA1	*TP53*	0	0	0.964	0	0.928	0.997
TP53	*RB1*	0	0	0.672	0.900	0.926	0.997
PCGF4	*PHC1*	0	0	0.974	0	0.910	0.997
UBE3A	*TP53*	0	0	0.982	0	0.818	0.996
SOX2	*PAX6*	0	0	0.964	0	0.908	0.996
SMAD2	*CREBBP*	0	0	0.951	0.800	0.486	0.994
GJA1	*MAPK3*	0	0	0.672	0.800	0.900	0.993
SMAD2	*MAPK3*	0	0	0.672	0.800	0.900	0.993
BRCA1	*RB1*	0	0	0.892	0	0.935	0.992
CASP3	*RB1*	0	0	0.897	0.900	0.282	0.992
RB1	*CDKN1B*	0	0	0	0.900	0.915	0.991
PTPN11	*EGFR*	0	0	0.900	0	0.909	0.99
AKT1	*CDKN1B*	0	0	0.892	0	0.912	0.99
EGFR	*DCN*	0	0	0.892	0	0.900	0.989
APP	*TGFB1*	0	0	0.982	0	0.234	0.986
HMGB1	*TP53*	0	0	0.982	0	0.124	0.984
PPARBP	*TP53*	0	0	0.964	0	0.518	0.982
GLI2	*ZIC2*	0.666	0	0.672	0.800	0.724	0.981
EGFR	*MAPK3*	0	0	0	0.800	0.903	0.98
BCL2L1	*AKT1*	0	0	0	0.900	0.788	0.978
COL2A1	*DCN*	0	0	0.837	0	0.845	0.974
PAX6	*SIX3*	0	0	0.681	0	0.914	0.972
GLI3	*ZIC2*	0.657	0	0.672	0.800	0.574	0.972
BRCA1	*LMO4*	0	0	0.672	0	0.915	0.972
NF1	*RB1*	0	0	0.672	0	0.916	0.972
PAX3	*MSX1*	0	0	0.672	0	0.910	0.97
GTF2I	*SMAD2*	0	0	0.672	0	0.907	0.969
TP53	*MAPK3*	0	0	0.672	0	0.903	0.968
PTPN11	*MAPK3*	0	0	0.672	0	0.900	0.967
CTNNB1	*EGFR*	0	0	0.701	0.800	0.423	0.965
SMAD2	*TP53*	0	0	0.672	0	0.892	0.964
MSX1	*TP53*	0	0	0.730	0	0.859	0.961
YWHAE	*CDKN1B*	0	0	0.946	0	0	0.946

Gene 1	Gene 2	CONFIDENCE SCORES					Combined confidence score
		Homology analysis	Coexpression evidence	Experimental data	Pathway data	Textmining	
COL2A1	*TGFB1*	0	0	0.837	0	0.648	0.942
RXRA	*PPARBP*	0	0	0.672	0	0.816	0.939
LDB2	*LMO4*	0	0	0.843	0	0.611	0.938
CTNNB1	*CREBBP*	0	0	0.933	0	0.028	0.934
CTNNB1	*MET*	0	0	0.672	0.800	0	0.934
UBE3A	*MECP2*	0	0	0	0	0.933	0.933
CASP3	*AKT1*	0	0	0.672	0	0.798	0.933
TP53	*TWIST1*	0	0	0	0.900	0.314	0.931
GLI3	*PTCH*	0	0	0	0	0.926	0.926
CASP3	*APP*	0	0	0.672	0	0.773	0.925
GLI2	*PTCH*	0	0	0	0	0.924	0.924
AKT1	*RB1*	0	0	0	0.900	0.241	0.924
NF1	*TP53*	0	0	0	0	0.923	0.923
TP53	*CDKN1B*	0	0	0	0	0.920	0.92
MSX1	*BMP4*	0	0	0	0	0.920	0.92
FGFR2	*TWIST1*	0	0	0	0	0.919	0.919
FGFR1	*TWIST1*	0	0	0	0	0.919	0.919
IGF2	*VEGF*	0	0	0	0	0.918	0.918
VEGF	*FGFB*	0	0	0	0	0.918	0.918
MSX1	*CREBBP*	0	0	0.892	0	0.238	0.917
BMP4	*DLX5*	0	0	0	0	0.917	0.917
FGFR3	*TWIST1*	0	0	0	0	0.917	0.917
VEGF	*TP53*	0	0	0	0	0.916	0.916
IGF2	*TP53*	0	0	0	0	0.916	0.916
CDKN1B	*TGFB1*	0	0	0	0	0.916	0.916
VEGF	*BMP4*	0	0	0	0	0.916	0.916
PAX3	*MET*	0	0	0	0	0.916	0.916
MECP2	*DLX5*	0	0	0	0	0.914	0.914
VEGF	*EGFR*	0	0	0	0	0.912	0.912
EGFR	*TP53*	0	0	0	0	0.912	0.912
VEGF	*ARNT*	0	0	0	0	0.912	0.912
SMAD2	*BMP7*	0	0	0	0	0.911	0.911
FGFB	*TGFB1*	0	0	0	0	0.910	0.91
VEGF	*AKT1*	0	0	0	0	0.909	0.909
CASP3	*TP53*	0	0	0	0	0.907	0.907
MET	*TP53*	0	0	0	0	0.906	0.906
VEGF	*GJA1*	0	0	0	0	0.906	0.906
PAX3	*BMP4*	0	0	0	0	0.906	0.906
VEGF	*CXCR4*	0	0	0	0	0.906	0.906
CASP3	*BCL2L1*	0	0	0	0	0.906	0.906
VEGF	*TGFB1*	0	0	0	0	0.905	0.905
EGFR	*TGFB1*	0	0	0	0	0.903	0.903
VEGF	*MAPK3*	0	0	0	0	0.903	0.903
AKT1	*MAPK3*	0	0	0	0	0.902	0.902
FGFB	*CXCR4*	0	0	0	0	0.902	0.902
CTNNB1	*PPARBP*	0	0	0	0.900	0	0.900

Gene 1	Gene 2	CONFIDENCE SCORES					Combined confidence score
		Homology analysis	Coexpression evidence	Experimental data	Pathway data	Textmining	
VEGF	*CASP3*	0	0	0	0	0.900	0.900
PAX3	*BCL2L1*	0	0	0	0	0.900	0.900
SOX2	*SIX3*	0	0	0	0	0.900	0.900
GLI2	*AKT1*	0	0	0	0.900	0	0.900
NR2F2	*AKT1*	0	0	0	0.900	0	0.900
COL2A1	*FGFR3*	0	0	0	0	0.900	0.900
MAPK3	*NGFR*	0	0	0.672	0	0.697	0.900

Table B.3: The top-ranked genes had the following Gene Ontology terms annotations significantly over-represented among them.

Term	Count	*P*-value[1]
BIOLOGICAL PROCESS		
Development	53	4.5×10^{-23}
Organ development	28	2.9×10^{-17}
Regulation of cellular process	55	3.9×10^{-13}
Regulation of biological process	56	1.7×10^{-12}
Regulation of physiological process	53	2.5×10^{-12}
Morphogenesis	25	2.9×10^{-12}
Regulation of cellular physiological process	52	3.8×10^{-12}
Regulation of transcription DNA-dependent	39	2.3×10^{-11}
Transcription DNA-dependent	39	7.4×10^{-11}
Regulation of cellular metabolism	41	9.6×10^{-11}
Regulation of transcription	39	1.5×10^{-10}
Regulation of nucleic acid metabolism 39 44.8%	39	2.5×10^{-10}
Regulation of metabolism	41	3.2×10^{-10}
Skeletal development	12	3.9×10^{-10}
Transcription	39	6.9×10^{-10}
System development	20	1.0×10^{-9}
Organ morphogenesis	14	2.8×10^{-9}
Nervous system development	19	6.3×10^{-9}
Cell communication	44	5.1×10^{-8}
Signal transduction	41	1.3×10^{-7}
Transcription from RNA polymerase II promoter	17	8.3×10^{-7}
Cell differentiation	16	1.7×10^{-6}
Regulation of transcription from RNA polymerase II promoter	12	7.6×10^{-6}
Growth	10	1.1×10^{-5}
Positive regulation of cellular process	16	1.6×10^{-5}
Positive regulation of cellular metabolism	9	2.4×10^{-5}
Embryonic development	7	2.7×10^{-5}
Nucleic acid metabolism	40	3.2×10^{-5}
Positive regulation of cellular physiological process	14	3.9×10^{-5}
Positive regulation of metabolism	9	5.2×10^{-5}
Pattern specification	5	6.0×10^{-5}

Term	Count	*P*-value[1]
Positive regulation of physiological process	14	6.3×10^{-5}
Mesoderm development	5	7.2×10^{-5}
Enzyme linked receptor protein signalling pathway	9	8.9×10^{-5}
Positive regulation of biological process	16	8.9×10^{-5}
Positive regulation of transcription, DNA-dependant	7	1.3×10^{-4}
Brain development	5	1.4×10^{-4}
Negative regulation of cellular process	16	3.8×10^{-4}
Negative regulation of cellular physiological process	15	4.5×10^{-4}
Positive regulation of transcription	7	4.9×10^{-4}
Negative regulation of physiological process	15	5.5×10^{-4}
Cell proliferation	13	5.6×10^{-4}
Transmembrane receptor protein tyrosine kinase signalling pathway	7	5.8×10^{-4}
Positive regulation of nucleic acid metabolism	7	5.8×10^{-4}
Tissue development	7	6.5×10^{-4}
Negative regulation of biological process	16	8.1×10^{-4}
Regulation of progression through cell cycle	12	0.0011
Segmentation	3	0.0011
Regulation of cell cycle	12	0.0011
Regulation of cell size	7	0.0011
Cell growth	7	0.0011
Apoptosis	13	0.0012
Programmed cell death	13	0.0012
Central nervous system development	6	0.0013
Cell death	13	0.0016
Death	13	0.0018
Angiogenesis	5	0.002
Blood vessel morphogenesis	5	0.0024
Blood vessel development	5	0.0024
Vasculature development	5	0.0024
Embryonic pattern specification	3	0.0029
Neurophysiological process	11	0.0031
Cell surface receptor linked signal transduction	16	0.0034
Androgen receptor signalling pathway	4	0.0045
Regulation of apoptosis	9	0.0063
Regulation of programmed cell death	9	0.0064
Regulation of cell proliferation	8	0.0071
Steroid hormone receptor signalling pathway	4	0.0073
Cell cycle	13	0.0083
Intracellular receptor-mediated signalling pathway	4	0.0084
Negative regulation of progression through cell cycle	6	0.0097
Protein amino acid phosphorylation	11	0.01
Negative regulation of transcription, DNA-dependent	5	0.011
Cell motility	7	0.011
Locomotion	7	0.011
Localization of cell	7	0.011
Positive regulation of cell proliferation	5	0.015
Cellular morphogenesis	7	0.02

Term	Count	*P*-value[1]
Ossification	3	0.02
Biomineral formation	3	0.02
Periodic pa itioning	2	0.021
Segment polarity determination	2	0.021
Negative regulation of transcription from RNA polymerase II promoter	4	0.022
Bone remodeling	3	0.022
Tissue remodeling	3	0.022
Cell organization and biogenesis	19	0.023
Primary metabolism	57	0.025
Striated muscle development	3	0.026
Negative regulation of apoptosis	5	0.028
Negative regulation of programmed cell death	5	0.029
Muscle development	4	0.031
Positive regulation of transcription from RNA polymerase II promoter	3	0.037
Phosphorylation	11	0.038
Negative regulation of transcription	5	0.038
Blastoderm segmentation	2	0.041
Sensory perception of sound	4	0.043
Sensory perception of mechanical stimulus	4	0.043
Negative regulation of nucleic acid metabolism	5	0.046
Cell cycle checkpoint	3	0.047
Cell migration	4	0.049
Cell-cell signalling	8	0.049
Positive regulation of organismal physiological process	3	0.049
MOLECULAR FUNCTION		
Transcription regulator activity	36	4.6×10^{-15}
Transcription factor activity	29	1.4×10^{-13}
DNA binding	39	1.6×10^{-11}
Sequence-specific DNA binding	18	9.3×10^{-11}
Protein binding	59	1.7×10^{-9}
Signal transducer activity	37	8.3×10^{-9}
Transcription factor binding	15	1.1×10^{-7}
Nucleic acid binding	39	1.5×10^{-6}
Growth factor activity	9	2.4×10^{-6}
Transcription cofactor activity	12	4.9×10^{-6}
Transcriptional activator activity	11	8.2×10^{-6}
Binding	78	1.0×10^{-5}
Transcription coactivator activity	9	3.5×10^{-5}
Receptor binding	14	5.9×10^{-4}
Heparin binding	6	1.3×10^{-4}
Steroid hormone receptoractivity	5	4.1×10^{-4}
Ligand-dependent nuclear receptor activity	5	5.1×10^{-4}
Glycosaminoglycan binding	6	5.4×10^{-4}
Fibroblast growth factor receptor activity	3	5.7×10^{-4}
Polysaccharide binding	6	6.5×10^{-4}
Pattern binding	6	8.9×10^{-4}

Term	Count	*P*-value[1]
Receptor activity	18	0.0019
Tumor suppressor	3	0.0026
Protein-tyrosine kinase activity	7	0.0029
Obsolete molecular function	8	0.0046
Cell cycle regulator	3	0.006
Nuclear hormone receptor binding	4	0.0092
Hormone receptor binding	4	0.0092
Transmembrane receptor activity	11	0.011
Double-stranded DNA binding	3	0.013
Androgen receptor binding	3	0.016
Carbohydrate binding	6	0.017
RNA polymerase II transcription factor activity	6	0.018
Cytokine activity	5	0.018
Steroid hormone receptor binding	3	0.025
Zinc ion binding	20	0.028
Identical protein binding	5	0.041
Transforming growth factor beta receptor binding	2	0.048

[1]*P-value obtained using the original candidate gene list as a background list to the top-ranked candidate genes*

Table B.4: The promoter elements that have been found in the target promoter set relative to the background promoter set.

Promoter Elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
+1 SRF	4.9832	6.99	3.41	0.00010	0.00004	38	350	544	10255	0.067
-1 Nkx6-2	3.8034	7.17	3.6	0.00006	0.00003	39	369	544	10255	0.091
-1 IPF1	3.3286	5.51	2.98	0.00005	0.00003	30	306	544	10255	1.000
-1 SRF	3.1818	6.25	3.81	0.00009	0.00005	34	391	544	10255	1.000
+1 TBX5	3.0911	7.9	4.45	0.00007	0.00004	43	456	544	10255	0.406
+1 Pit-1	3.0386	5.33	3.15	0.00005	0.00003	29	323	544	10255	1.000
+1 HSF1	2.7461	8.64	5.03	0.00007	0.00005	47	516	544	10255	0.448
+1 HMG IY	2.7395	11.95	7.1	0.00010	0.00006	65	728	544	10255	0.060
+1 IRF1	2.6653	12.68	7.93	0.00012	0.00007	69	813	544	10255	0.147
+1 MEF-2	2.5417	7.9	5.02	0.00007	0.00004	43	515	544	10255	1.000
-1 Hand1:E47	2.4874	5.7	3.62	0.00005	0.00003	31	371	544	10255	1.000
+1 Cdx-2	2.3874	10.85	7.11	0.00010	0.00006	59	729	544	10255	1.000
+1 C/EBPgamma	2.3669	7.9	5.03	0.00007	0.00004	43	516	544	10255	1.000
-1 Sp3	2.2802	6.25	4.08	0.00005	0.00004	34	418	544	10255	1.000
-1 Pbx-1	2.2091	15.07	10.13	0.00014	0.00009	82	1039	544	10255	0.315
-1 DBP	2.1563	9.93	6.57	0.00008	0.00006	54	674	544	10255	1.000
+1 C/EBPdelta	2.1086	5.15	3.52	0.00004	0.00003	28	361	544	10255	1.000
+1 NF-AT	2.0871	5.7	4.16	0.00006	0.00004	31	427	544	10255	1.000
-1 TBP	2.0639	9.38	6.49	0.00008	0.00006	51	666	544	10255	1.000

Table B.5: Pairs of promoter elements at maximum mutual distance of 50 nucleotides that have been found in the target promoter set relative to the background promoter set.

Pairs of promoter elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
-1 LBP-1/+1 AP-2	10.7173	5.15	1.64	0.00006	0.00002	28	168	544	10255	0.617500
+1 Pax-4/+1 FAC1	9.354	5.33	1.84	0.00006	0.00002	29	189	544	10255	1.000000
-1 Spz1/-1 ETF	8.8907	6.43	2.71	0.00014	0.00004	35	278	544	10255	1.000000
-1 MZF1/-1 VDR	8.3078	7.9	3.15	0.00017	0.00005	43	323	544	10255	0.193900
+1 FAC1/+1 Pax-4	7.8354	5.15	2.3	0.00010	0.00003	28	236	544	10255	1.000000
-1 MAZ/+1 ZF5	7.7859	6.25	2.04	0.00006	0.00002	34	209	544	10255	0.062440
+1 Pax-4/-1 MZF1	7.6641	5.51	2.4	0.00008	0.00003	30	246	544	10255	1.000000
-1 ZF5/-1 MAZ	7.2777	6.62	2.94	0.00011	0.00004	36	301	544	10255	1.000000
+1 ETF/-1 Sp1	6.6357	6.8	4.17	0.00037	0.00009	37	428	544	10255	1.000000
-1 ZF5/+1 C/EBP	6.2165	5.15	2.11	0.00006	0.00002	28	216	544	10255	1.000000
-1 MAZ/+1 AP-2	5.9377	5.7	2.38	0.00008	0.00003	31	244	544	10255	1.000000
+1 C/EBP/+1 Spz1	5.8262	6.07	2.44	0.00006	0.00003	33	250	544	10255	1.000000
+1 Sp1/-1 Spz1	5.7749	10.85	5.45	0.00030	0.00010	59	559	544	10255	1.000000
+1 ZF5/+1 MAZ	5.3521	6.8	2.78	0.00008	0.00004	37	285	544	10255	1.000000
+1 ZF5/-1 MAZ	5.171	5.33	2.15	0.00005	0.00003	29	220	544	10255	1.000000
-1 VDR/-1 MAZ	5.1338	6.8	3.17	0.00014	0.00006	37	325	544	10255	1.000000
+1 ZF5/-1 AP-2gamma	4.9973	5.33	2.33	0.00005	0.00002	29	239	544	10255	1.000000
+1 E2F/-1 VDR	4.9316	7.17	4.21	0.00015	0.00005	39	432	544	10255	1.000000
+1 Spz1/+1 Spz1	4.903	8.82	5.14	0.00021	0.00007	48	527	544	10255	1.000000
-1 Spz1/-1 MAZ	4.8704	5.7	2.72	0.00009	0.00004	31	279	544	10255	1.000000
+1 Pax-4/-1 Spz1	4.8571	9.93	4.85	0.00012	0.00005	54	497	544	10255	1.000000
-1 VDR/+1 VDR	4.8498	6.62	3.21	0.00009	0.00004	36	329	544	10255	1.000000
+1 Oct-1/+1 Pax-4	4.706	6.43	2.96	0.00008	0.00003	35	304	544	10255	1.000000
+1 Spz1/-1 Pax-4	4.6875	9.01	4.85	0.00012	0.00005	49	497	544	10255	1.000000
-1 ETF/+1 E2F	4.5575	6.8	3.66	0.00011	0.00005	37	375	544	10255	1.000000
-1 MAZ/-1 Spz1	4.5301	5.33	2.37	0.00007	0.00004	29	243	544	10255	1.000000
-1 Spz1/+1 VDR	4.5179	6.07	2.75	0.00007	0.00003	33	282	544	10255	1.000000
-1 MZF1/-1 ZF5	4.4631	5.7	2.79	0.00007	0.00003	31	286	544	10255	1.000000
-1 Spz1/+1 Pax-4	4.4298	9.01	4.85	0.00012	0.00005	49	497	544	10255	1.000000
+1 Spz1/+1 VDR	4.3932	11.76	6.92	0.00027	0.00010	64	710	544	10255	1.000000
+1 ETF/-1 ZF5	4.2011	9.93	6.5	0.00029	0.00011	54	667	544	10255	1.000000
-1 Oct-1/+1 Pax-4	4.1981	6.8	3.16	0.00007	0.00004	37	324	544	10255	1.000000
+1 ETF/+1 Sp1	4.1285	8.09	5.8	0.00061	0.00021	44	595	544	10255	1.000000
-1 VDR/+1 Pax-4	4.1164	10.29	5.88	0.00018	0.00008	56	603	544	10255	1.000000
-1 Pax-4/-1 Pax-2	4.0956	5.15	2.55	0.00005	0.00002	28	262	544	10255	1.000000
+1 AP-2/-1 ETF	4.0559	9.19	5.3	0.00023	0.00010	50	544	544	10255	1.000000
-1 C/EBP/-1 GEN_INI	4.0546	6.99	3.43	0.00014	0.00007	38	352	544	10255	1.000000
-1 Pax-4/-1 Oct-1	4.0195	6.07	3.11	0.00007	0.00004	33	319	544	10255	1.000000
-1 Tst-1/+1 Pax-4	4.0172	5.33	2.54	0.00005	0.00002	29	260	544	10255	1.000000
-1 ETF/+1 ZF5	3.8938	9.38	5.66	0.00018	0.00008	51	580	544	10255	1.000000
-1 ETF/-1 ZF5	3.7955	11.03	6.13	0.00022	0.00011	60	629	544	10255	1.000000
-1 C/EBP/-1 Pax-4	3.6539	9.38	4.84	0.00009	0.00005	51	496	544	10255	1.000000
+1 ETF/+1 AP-2	3.6306	9.19	6.51	0.00030	0.00011	50	668	544	10255	1.000000

Pairs of promoter elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
+1 C/EBP/+1 GEN_INI	3.5705	7.54	3.92	0.00016	0.00009	41	402	544	10255	1.000000
-1 C/EBP/-1 Oct-1	3.5228	5.51	3.46	0.00009	0.00004	30	355	544	10255	1.000000
+1 ZF5/-1 ETF	3.4419	10.85	5.69	0.00016	0.00009	59	583	544	10255	1.000000
-1 Spz1/-1 AP-2	3.4355	8.27	5.38	0.00017	0.00008	45	552	544	10255	1.000000
-1 VDR/-1 MZF1	3.4345	5.88	3.22	0.00009	0.00005	32	330	544	10255	1.000000
-1 VDR/-1 VDR	3.2937	15.07	8.08	0.00028	0.00016	82	829	544	10255	0.129800
+1 ETF/+1 ETF	3.2705	10.85	8.06	0.00062	0.00025	59	827	544	10255	1.000000
-1 E2F/-1 ETF	3.2424	7.72	5.54	0.00022	0.00009	42	568	544	10255	1.000000
+1 ZF5/-1 VDR	3.2332	9.93	6.35	0.00017	0.00008	54	651	544	10255	1.000000
+1 VDR/+1 MAZ	3.2282	6.43	2.83	0.00006	0.00005	35	290	544	10255	1.000000
-1 VDR/-1 E2F	3.211	9.38	6.46	0.00021	0.00009	51	662	544	10255	1.000000
-1 MAZ/-1 MAZ	3.1829	5.15	1.49	0.00006	0.00006	28	153	544	10255	0.113400
-1 MZF1/+1 Sp1	3.0484	5.7	2.37	0.00006	0.00005	31	243	544	10255	1.000000
+1 Sp1/-1 VDR	3.0235	11.76	7.01	0.00028	0.00016	64	719	544	10255	1.000000
-1 Oct-1/-1 Pax-4	3.0033	6.07	3.43	0.00007	0.00004	33	352	544	10255	1.000000
+1 AP-2/+1 GEN_INI	2.9689	6.62	3.68	0.00015	0.00009	36	377	544	10255	1.000000
-1 ZF5/+1 ETF	2.9515	9.56	7.01	0.00024	0.00011	52	719	544	10255	1.000000
-1 ZF5/-1 ETF	2.916	12.5	7.22	0.00022	0.00013	68	740	544	10255	1.000000
-1 Spz1/+1 Sp1	2.8755	10.48	6.07	0.00019	0.00011	57	622	544	10255	1.000000
+1 Sp1/-1 MAZ	2.8744	5.7	2.76	0.00009	0.00006	31	283	544	10255	1.000000
-1 E2F/+1 ETF	2.8346	5.15	4.59	0.00016	0.00006	28	471	544	10255	1.000000
-1 Sp1/-1 MAZ	2.827	5.88	2.94	0.00013	0.00010	32	301	544	10255	1.000000
-1 VDR/+1 E2F	2.7991	8.09	4.87	0.00011	0.00007	44	499	544	10255	1.000000
-1 Spz1/-1 Sp1	2.7847	12.32	6.63	0.00031	0.00021	67	680	544	10255	1.000000
-1 C/EBP/+1 Oct-1	2.7834	6.07	3.59	0.00008	0.00005	33	368	544	10255	1.000000
-1 Spz1/-1 E2F	2.7777	8.46	5.56	0.00012	0.00006	46	570	544	10255	1.000000
-1 MAZ/-1 Sp1	2.7765	5.51	2.96	0.00014	0.00009	30	304	544	10255	1.000000
-1 Pax-4/+1 Oct-1	2.7422	5.33	3.25	0.00006	0.00004	29	333	544	10255	1.000000
-1 E2F-1/-1 Pax-4	2.7262	6.8	4.85	0.00011	0.00006	37	497	544	10255	1.000000
-1 ZF5/-1 VDR	2.7201	13.42	8.45	0.00022	0.00013	73	867	544	10255	1.000000
-1 Spz1/+1 ZF5	2.7194	9.38	5.53	0.00010	0.00006	51	567	544	10255	1.000000
-1 VDR/+1 AP-2	2.7174	12.13	7.86	0.00023	0.00013	66	806	544	10255	1.000000
-1 Spz1/+1 AP-2	2.6995	8.27	6.1	0.00017	0.00008	45	626	544	10255	1.000000
+1 VDR/-1 E2F-1	2.636	5.33	3.97	0.00011	0.00006	29	407	544	10255	1.000000
-1 Pax-4/+1 GEN_INI	2.6326	8.82	4.98	0.00016	0.00011	48	511	544	10255	1.000000
+1 E2F/+1 ETF	2.6312	11.4	7.67	0.00023	0.00013	62	787	544	10255	1.000000
+1 VDR/-1 AP-2	2.5728	10.48	7.78	0.00023	0.00012	57	798	544	10255	1.000000
+1 C/EBP/-1 C/EBP	2.5384	8.09	5.33	0.00011	0.00006	44	547	544	10255	1.000000
+1 ETF/+1 E2F	2.4797	9.19	7.35	0.00024	0.00012	50	754	544	10255	1.000000
-1 Pax-4/+1 Spz1	2.4614	7.35	5.84	0.00012	0.00006	40	599	544	10255	1.000000
+1 CDX/+1 CDX	2.433	5.15	2.08	0.00005	0.00005	28	213	544	10255	1.000000
-1 Pax-4/+1 VDR	2.4297	8.64	6.38	0.00014	0.00008	47	654	544	10255	1.000000
+1 GC box/+1 Spz1	2.397	5.33	3.68	0.00007	0.00004	29	377	544	10255	1.000000
+1 C/EBP/+1 C/EBP	2.3729	7.54	4.87	0.00009	0.00006	41	499	544	10255	1.000000
-1 ETF/-1 ETF	2.3364	6.62	5.26	0.00036	0.00019	36	539	544	10255	1.000000
-1 AP-2/-1 Pax-4	2.3118	8.82	6.24	0.00013	0.00008	48	640	544	10255	1.000000

Pairs of promoter elements	ORI	TAR (%)	BCG (%)	Probability of finding PE in target set	Probability of finding PE in background set	TAR (n)	BCG (n)	TAR Total	BCG Total	*P*-value
-1 Pax-4/-1 ZF5	2.3093	9.38	6.32	0.00011	0.00007	51	648	544	10255	1.000000
+1 ZF5/+1 Spz1	2.3087	10.48	7.54	0.00015	0.00009	57	773	544	10255	1.000000
-1 AP-2/+1 AP-2	2.2735	17.1	12.66	0.00043	0.00026	93	1298	544	10255	1.000000
-1 E2F-1/+1 Sp1	2.2487	8.09	5.81	0.00030	0.00019	44	596	544	10255	1.000000
+1 C/EBP/-1 Pax-4	2.2468	8.09	5.1	0.00008	0.00005	44	523	544	10255	1.000000
-1 C/EBP/-1 C/EBP	2.2411	7.54	4.79	0.00008	0.00006	41	491	544	10255	1.000000
-1 E2F/+1 AP-2	2.2048	13.6	9.97	0.00028	0.00017	74	1022	544	10255	1.000000
+1 E2F/-1 ETF	2.1983	5.51	3.8	0.00007	0.00005	30	390	544	10255	1.000000
+1 AP-2/+1 AP-2	2.1656	15.99	12.96	0.00048	0.00027	87	1329	544	10255	1.000000
-1 AP-2/-1 AP-2	2.1535	14.52	12.75	0.00055	0.00029	79	1308	544	10255	1.000000
+1 ZF5/+1 ETF	2.1293	12.32	9.63	0.00029	0.00018	67	988	544	10255	1.000000
+1 AP-2/+1 E2F	2.1068	15.62	11.86	0.00028	0.00018	85	1216	544	10255	1.000000
+1 E2F/-1 GC box	2.0944	5.51	3.52	0.00006	0.00005	30	361	544	10255	1.000000
-1 Pax-4/-1 C/EBP	2.0906	7.9	5.62	0.00009	0.00006	43	576	544	10255	1.000000
+1 ETF/+1 ZF5	2.0791	8.46	9.01	0.00033	0.00015	46	924	544	10255	1.000000
+1 AP-2/+1 ZF5	2.0762	20.22	15.76	0.00045	0.00028	110	1616	544	10255	1.000000
-1 ZF5/-1 Spz1	2.0761	7.54	6.39	0.00014	0.00008	41	655	544	10255	1.000000
+1 Pax-4/-1 C/EBP	2.071	8.27	5.61	0.00009	0.00006	45	575	544	10255	1.000000
-1 ZF5/+1 AP-2	2.0443	17.1	14.69	0.00048	0.00027	93	1506	544	10255	1.000000
-1 Spz1/+1 GEN_INI	2.0393	5.7	3.77	0.00011	0.00008	31	387	544	10255	1.000000
+1 VDR/+1 AP-2	2.0253	9.93	7.66	0.00017	0.00011	54	786	544	10255	1.000000
+1 ETF/-1 AP-2	2.0252	8.46	7.29	0.00024	0.00014	46	748	544	10255	1.000000
+1 AP-2/-1 AP-2	2.0129	17.1	13.32	0.00040	0.00025	93	1366	544	10255	1.000000
+1 ZF5/-1 Pax-4	2.0116	10.48	8.16	0.00015	0.00010	57	837	544	10255	1.000000
+1 VDR/+1 Spz1	2.0033	7.9	7.03	0.00019	0.00011	43	721	544	10255	1.000000
+1 Pax-4/+1 VDR	2.0013	6.07	4.71	0.00008	0.00005	33	483	544	10255	1.000000

Table B.6: This table displays the intersect of the top-ranked candidate genes for XLMR and linked regions for XLMR of unknown genetic aetiology. The shaded blocks indicate where a prioritized candidate gene falls within a linked region, and is therefore a putative candidate for that syndrome. The numbers at the top of the table correlate to the numbering used in Table 3.2 and Figure 3.1.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
		P ARM									Q ARM																		SPANNING				
9	**ATRX**										x	x	x	x	x														x		x	x	x
9	**EFNB1**										x	x	x	x															x	x	x	x	x
9	**NONO**										x	x	x	x															x	x	x	x	x
9	**PDZD11**										x	x	x	x															x	x	x	x	x
9	**TAF9B**										x	x	x	x	x														x		x	x	x
7	CXorf1																			x	x	x		x	x	x	x						
6	MAGED1							x		x																			x	x	x		x
6	MAOA							x	x																				x	x	x		x
6	PCTK1							x	x																				x	x	x		x
6	TMEM29							x		x																			x	x	x		x
6	TMEM29B							x		x																			x	x	x		x
6	UBE1							x	x																				x	x	x		x
6	USP11							x	x																				x	x	x		x
5	ARAF							x																					x	x	x		x
5	BGN																							x	x	x	x	x					
5	FAM39DP																							x	x	x	x	x					
5	FAM39E																							x	x	x	x	x					
5	FLNA																							x	x	x	x	x					
5	FUNDC2																							x	x	x	x	x					
5	HDAC6							x																					x	x	x		x
5	IDS																							x	x	x	x	x					

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
		P ARM									Q ARM																		SPANNING				
5	L1CAM																							x	x	x	x	x					
5	MECP2																							x	x	x	x	x					
5	MID1IP1							x	x																				x	x			x
5	MORF4L2													x		x													x		x		x
5	NGFRAP1													x		x													x		x		x
5	PLP1													x		x													x		x		x
5	PQBP1							x																					x	x	x		x
5	SLC6A8																							x	x	x	x	x					
5	SYN1							x																					x	x	x		x
5	TIMP1							x																					x	x	x		x
5	TSPAN7							x	x																				x	x			x
5	WDR13							x																					x	x	x		x
4	EIF2S3		x		x	x	x																										
4	FGF13																			x	x	x	x										
4	FHL1																			x	x	x	x										
4	HPRT1																			x	x	x	x										
4	HTATSF1																			x	x	x	x										
4	MAGED2																												x	x	x		x
4	PLS3															x		x													x		x
4	TRO																												x	x	x		x
3	CUL4B																	x	x														x
3	LAMP2																	x	x														x
3	PGRMC1																	x	x														x
3	SFRS17A	x				x	x																										
2	AP1S2					x	x																										
2	DMD				x																									x			

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
		P ARM									Q ARM																		SPANNING				
2	GPM6B					x	x																										
2	MSL3L1					x	x																										
2	RBBP7					x	x																										
1	AIFM1																			x													
1	ZDHHC9																			x													

Addendum C :

Reagents & Equipment – Suppliers

REAGENTS

REAGENT	SUPPLIER
1 Kb Plus DNA ladder (10787-018)	Invitrogen Corporation, CA United States
Agarose, (Molecular grade)	Whitehead Scientific, Brackenfell South Africa
AmpliTaq Gold Taq polymerase, buffer & MgCl	Applied Biosystems, NJ United States
BigDye® terminator v3.1 cycle sequencing kit	Applied Biosystems, NJ United States
dNTPs	Bioline, London UK
Ethidium Bromide aqueous solution	Sigma Aldrich, MO United States
Flexigene DNA kit	Qiagen Inc, CA United States
SEQ96 sequencing reaction cleanup system	Montage™ Millipore, MA United States
MultiScreen® $PCR_{\mu 96}$ plate	Millipore, MA United States
Primers	Inqaba Biotec, Gauteng SA

EQUIPMENT

INSTRUMENT	SUPPLIER
ABI 3130 genetic analyser	Applied Biosystems, NJ United States
Beckman CS-6R centrifuge	Beckman Coultier, CA United States
Eppendorf® Mastercycler® Gradient PCR machine	Eppendorf, Hamburg Germany
Geneamp® PCR system 2720	Applied Biosystems, NJ United States
Maxi Electrophoresis Unit	Sigma Aldrich, MO United States
Millipore MilliVac® Maxi vacuum manifold	Millipore, MA United States
Nanodrop® ND-1000 Spectrophotometer	Thermo Fisher Scientific, MA United States

Addendum D :

Ethics Approval Certificates