

Genetic Variants Experiencing Sexually Antagonistic Selection in Humans

Jinghua FENG

March 2012

A thesis submitted for the degree of Doctor of Philosophy
of the Australian National University

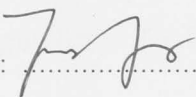


Australian
National
University

Declaration

The thesis comprises only my original work towards the PhD in The Australian National University unless otherwise referenced. The document has not been submitted for qualifications at any other academic institution.

Name: Jinghua Feng

Signature: 

Date: 27/3/2012

Acknowledgements

Heartfelt appreciation goes to my supervisors, Associate Professor Gavin Huttley and Professor Simon Easteal. Their knowledge and assistance during my research is invaluable to me.

I thank Dr. Matthew Wakefield and Dr. Rohan Williams from my supervisory panel for their advice and support.

I am grateful to Dr Klara Verbyla for her careful proofreading and advice on writing thesis.

I thank JCSMR for the scholarship support. Especially, my appreciation goes to Dr. Anna Cowan for her work for students in the school.

Lastly, I greatly appreciate my wife Dan for her understanding and support.

Abstract

The two sexes are so different that a genetic variant may function differentially in males and females. Because the evolutionary interests of the sexes never exactly coincide and the two sexes share most of their genetic materials (all except the Y chromosome in mammals), conflicting natural selection can arise. Shared variants or traits undergo differential natural selection between the sexes, resulting in selection for a shared trait in one sex displacing the other sex from its optimum. The phenomenon of sexually antagonistic (SA) selection has been demonstrated in a wide range of taxa, including animals, insects and dioecious plants. However, it has not been widely studied in humans, and clear evidence is still rare. Furthermore, very little is known for the genetics underlying SA selection.

In this thesis, SA selection in humans is studied at the viability level. Males and females are expected to share the same genotype distribution at autosomal loci. Thus, significant differences in genotype distribution at single-nucleotide polymorphisms (SNPs) between the sexes is an indicator of SA selection at the viability level. Genome-wide SNP data were the main materials used in the study. SNP data was obtained from white European control cohort of the Wellcome

Trust Case Control Consortium, which is a representative population sample. After quality control, $\sim 4,800$ individuals were included in the analysis with approximately equal numbers of males and females, and for each individual there were ~ 1.2 million genotyped autosomal SNPs available.

Comparison of genotype distribution between the sexes at these SNPs provides evidence for ongoing conflicting natural selection between the sexes in humans. This is confirmed by examination of SNPs from genes associated with early onset disease and from genes related to sex hormone function. The former indicates that it is sex-specific differences of viability and/or mortality that lead to differences in genotype distributions between the sexes. The latter indicates that differences in sex hormones contribute to sex-differential viability or mortality.

Candidate genes from regions with signatures of SA selection are presented. These candidate genes are important materials for studying the genetic mechanisms underlying SA selection. They may be associated with sex-differential prenatal viability and thus with spontaneous abortion. They may also be associated with sex differences in infant health and disease occurrence later in life.

Contents

Declaration	i
Acknowledgements	iii
Abstract	v
Contents	vii
List of Figures	xi
List of Tables	xv
Abbreviations	xvii
1 Introduction	1
1.1 Conflict between the sexes	1
1.2 Reported evidence for sexual antagonism	4
1.3 Aims of the study	7
2 Genomic data and quality control	9

- 2.1 SNP data 10
 - 2.1.1 Subjects 10
 - 2.1.2 The related genotyping technologies 11
- 2.2 Data quality control 13
 - 2.2.1 Importance of quality control 13
 - 2.2.2 Quality control on individual level 14
 - 2.2.3 Quality control at SNP level 17
 - 2.2.3.1 Exclusion of the non-autosomal SNPs 17
 - 2.2.3.2 Common QC filters from the WTCCC 18
 - 2.2.3.3 QC against SNPs in segmental genome duplications 21
 - 2.2.3.4 QC for the combination of control samples 24
 - 2.2.3.5 Combination of SNP data from the two genotyping platforms 24
- 2.3 Summary 25
- 3 Evidence for SA selection in humans 29**
 - 3.1 Statistical methods 30
 - 3.1.1 Huttley and Wilson’s log-linear models 30
 - 3.1.2 Adjustments for multiple testing 32
 - 3.1.2.1 Q value for false discovery rate 34
 - 3.1.2.2 The simpleM correction 34
 - 3.1.2.3 The permutation test 35
 - 3.2 Results 37

3.2.1	Examination of reported variants with sex differences in genotype distribution	37
3.2.2	Evidence for SA selection revealed by the genome-wide scan	42
3.2.2.1	The positive results from the genome-wide scan .	42
3.2.2.2	The validity of the signals for SA selection	51
3.3	Discussion	56
4	Characteristics of SA variant	61
4.1	Human diseases underlying SA selection	62
4.1.1	Sex-effects of diseases underlying SA selection	63
4.1.2	Signal concentration of diseases genes	66
4.2	Functional category enrichment	71
4.3	No enrichment in sex-hormone receptor binding loci	76
4.4	No enrichment in imprinted genes	80
4.5	No enrichment in regions of positively selected selection	83
4.6	The allelic effect underlying SA selection	85
4.7	Candidate genes	90
4.8	Summary	90
5	Conclusions	97
5.1	Discoveries from the study	98
5.2	Importance of the discoveries	99
5.3	Further study	101

Appendices	103
Bibliography	115

List of Figures

2.1	Structure of the human male sex chromosomes and the localization of PAR1 and PAR2.	17
2.2	The significant discordance in the genotype distribution between the sexes at the SNPs close to the pseudoautosomal boundary . .	19
2.3	The incomplete hybridization between the probe and the genomic sequences	23
3.1	Calculating the false discovery rate for multiple testing correction.	35
3.2	The distribution of signals for SA selection across a 1 Mb genomic region centered on BRCA2.	39
3.3	Signals for sex differences of genotype distribution in the genomic region with central rs1800907.	41
3.4	Quantile-quantile plot for the test statistics from the genome-wide scan.	44
3.5	Genome-wide picture of the signals for SA selection.	45

3.6	The eight genomic regions that reached $P < 1E-5$ in the genome-wide scan.	48
3.7	Comparison between the signals from Affymetrix V6.0 and Illumina 1.2M platforms	58
4.1	Comparisons of genotype distribution between cases and controls in each sex.	65
4.2	Q-Q plot for signals for SA selection from OMIM disease genes . .	67
4.3	Q-Q plot for signals for SA selection from hOMIM disease genes. .	68
4.4	Q-Q plot for signals for SA selection from all known autosomal genes.	70
4.5	Q-Q plots for the signals for SA selection from SNPs located in the disease genes from each of the four categories of disease-onset age.	72
4.6	Comparison of the signals for SA selection from SNPs located in the hOMIM disease genes from four categories of disease-onset age.	73
4.7	Q-Q plot for the 1,961 SNPs located in the 77 genes regarding sex hormone.	76
4.8	Q-Q plots for signals for SA selection from sex-hormone receptor binding loci.	79
4.9	Q-Q plot for the SNPs located inside the 55 imprinted genes, testing differences in genotype frequencies.	82
4.10	Q-Q plot for the SNPs located inside the 55 imprinted genes, testing differences in allelic frequencies.	83

4.11 Q-Q plot for the 92,557 SNPs located in the positively selected genomic region. 85

4.12 Q-Q plot for the comparison between the test statistics for sexually allelic differences (τM_A) and the expected values. 86

4.13 Q-Q plot for the comparisons between values of τM_A and of τM_{AA} , for all the SNPs used in genome-wide scan. 88

4.14 Q-Q plot for the comparisons between values of τM_A and of τM_{AA} , for the SNPs potentially experiencing SA selection. 89

List of Tables

2.1	Composition of each control and case cohort after quality control	17
2.2	Number of SNPs failing each QC filter	28
2.3	Proportion ratios of positives between filtered and unfiltered SNPs	28
3.1	Genotype distributions of rs1801406 in BRCA2	38
3.2	Observed and expected genotype frequencies at rs1801406 in BRCA2	40
3.3	Eight independent SNPs with the strongest signal ($P < 1E-5$) for SA selection revealed by genome-wide scan	46
3.4	Summary of candidate genes from the eight genomic regions showing P $< 1E-5$ for SA selection	52
3.5	Surrounding SNPs of the peak from the eight genomic regions showing $P < 1E-5$ for SA selection	53
3.6	P values at the eight SNPs with the strongest signals for the two control cohorts	54

4.1	Number of SNPs located within the hOMIM [16] disease genes of four categories of onset age	66
4.2	Enriched PANTHER biological process categories	74
4.3	Proportion of positive SNPs for SA selection at the estrogen and androgen receptor binding loci	78
4.4	Candidates genes selected from the gene groups enriched of signals for SA selection.	93
A1	Positive proportions for all the 242 PANTHER biological process categories.	103
A2	The 14 terms of biological process from the Gene Ontology regard- ing sex hormone	110
A3	Description for the selected imprinted genes	112

Abbreviations

58C	The 1958 Birth cohort
bp	Base pair
kb	Kilo base pairs
BRCA2	Breast cancer 2, early onset
FDR	False discovery rate
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
LD	Linkage disequilibrium
MAF	Minor allele frequency
OMIM	Online Mendelian inheritance in man database
PAR	Pseudoautosomal regions
PCR	Polymerase chain reaction
QC	Quality control
Q-Q	Quantile-quantile plot
SA	Sexually antagonistic
SHRBL	Sex-hormone receptor binding loci
SNP	Single-nucleotide polymorphism
UKBS	The UK Blood Services cohort
WTCCC	the Wellcome Trust Case-Control Consortium

1.1	Abbreviations
1.2	Abbreviations
1.3	Abbreviations
1.4	Abbreviations
1.5	Abbreviations
1.6	Abbreviations
1.7	Abbreviations
1.8	Abbreviations
1.9	Abbreviations
1.10	Abbreviations
1.11	Abbreviations
1.12	Abbreviations
1.13	Abbreviations
1.14	Abbreviations
1.15	Abbreviations
1.16	Abbreviations
1.17	Abbreviations
1.18	Abbreviations
1.19	Abbreviations
1.20	Abbreviations
1.21	Abbreviations
1.22	Abbreviations
1.23	Abbreviations
1.24	Abbreviations
1.25	Abbreviations
1.26	Abbreviations
1.27	Abbreviations
1.28	Abbreviations
1.29	Abbreviations
1.30	Abbreviations
1.31	Abbreviations
1.32	Abbreviations
1.33	Abbreviations
1.34	Abbreviations
1.35	Abbreviations
1.36	Abbreviations
1.37	Abbreviations
1.38	Abbreviations
1.39	Abbreviations
1.40	Abbreviations
1.41	Abbreviations
1.42	Abbreviations
1.43	Abbreviations
1.44	Abbreviations
1.45	Abbreviations
1.46	Abbreviations
1.47	Abbreviations
1.48	Abbreviations
1.49	Abbreviations
1.50	Abbreviations
1.51	Abbreviations
1.52	Abbreviations
1.53	Abbreviations
1.54	Abbreviations
1.55	Abbreviations
1.56	Abbreviations
1.57	Abbreviations
1.58	Abbreviations
1.59	Abbreviations
1.60	Abbreviations
1.61	Abbreviations
1.62	Abbreviations
1.63	Abbreviations
1.64	Abbreviations
1.65	Abbreviations
1.66	Abbreviations
1.67	Abbreviations
1.68	Abbreviations
1.69	Abbreviations
1.70	Abbreviations
1.71	Abbreviations
1.72	Abbreviations
1.73	Abbreviations
1.74	Abbreviations
1.75	Abbreviations
1.76	Abbreviations
1.77	Abbreviations
1.78	Abbreviations
1.79	Abbreviations
1.80	Abbreviations
1.81	Abbreviations
1.82	Abbreviations
1.83	Abbreviations
1.84	Abbreviations
1.85	Abbreviations
1.86	Abbreviations
1.87	Abbreviations
1.88	Abbreviations
1.89	Abbreviations
1.90	Abbreviations
1.91	Abbreviations
1.92	Abbreviations
1.93	Abbreviations
1.94	Abbreviations
1.95	Abbreviations
1.96	Abbreviations
1.97	Abbreviations
1.98	Abbreviations
1.99	Abbreviations
1.100	Abbreviations

Chapter 1

Introduction

1.1 Conflict between the sexes

A genetic variant can be beneficial to one sex but harmful to the other sex. It is well known that the two sexes differ in many ways in appearance, physique, anatomy, physiology, behaviour and reproductive role. In humans, males and females also often differ in social roles. From the perspective of a gene, sex can be considered an ‘environmental’ variable which differs between boys and girls (in childhood), and between men and women (in adulthood). Staying in such distinctive environments, a genetic variant shared by the two sexes can function differentially between the sexes, resulting in differential fitness effects [18, 161]. The shared variant may also work in the same direction between the sexes but with differential magnitudes, such as a variant causing a disease in both sexes with differences in prevalence or severity. From an evolutionary point of view,

these shared variants undergo differential or opposite natural selection between the sexes.

Sexually antagonistic (SA) selection occurs when the two sexes have different selective optima for their shared traits [18, 161]. It arises because the shared traits have a common genetic basis but undergo differential selection between the sexes [119]. Thus selection on one sex may cause the displacement of the other sex from its phenotypic optimum, reducing its fitness. The effects of SA selection can be described as intra- or interlocus conflict, depending on whether the target of selection is at a single locus or at more than one interacting loci [7]. An example of intralocus sexual conflict is that a gene is expressed in both sexes, but beneficial to one sex and detrimental to the other sex. As for interlocus sexual conflict, consider a case of mating decision described by Arnqvist and Rowe [7], where males pursue a high mating rate, while females prefer low mating rate for their best interests, and mating rate is determined by locus A in males and locus B in females. Alleles expressed at locus A, which can increase male mating rate, will spread in the population. The spread of these alleles at locus A is detrimental to females, which in turn prompts selection at locus B for alleles that cause a decrease in female mating rate. Spread of such alleles at locus B can then impact selection among alleles at locus A. Interlocus sexual conflict is thus mediated by the interaction between loci A and B in the two sexes. In this study, however, only intralocus sexual conflict will be considered.

By definition, SA selection generally produces unequal allele frequencies in

males and females each generation. However, single-locus models show that SA selection can maintain stable polymorphism with the balance of opposing forces on the two sexes [82, 136]. How allele frequencies change over generations at a single locus experiencing SA selection depends on the net effect between the differential selection acting on the two sexes. Simply based on whether selection acts in the same direction between the two sexes, two forms of allele frequency change can be predicted. The first is driven by selection acting in opposite directions between the sexes. In this tug of war, the selection force in the winner between the two sexes will determine how the allele frequency changes. Neutralization of opposing selection can maintain the balanced allelic frequencies. The second form of allele frequency change is directional, when natural selection is acting in the same direction but with differential magnitudes between the sexes. In that scenario, conflict between the sexes may be not as intense as the first one, but selection on one sex still slows adaptive evolution in the other sex [18]. Thus both of the two forms for sexual antagonism impede or slow the change of allele frequencies in each sex, and many see sexual antagonism as a brake on sexual coevolution [30, 45, 45, 88, 124, 140].

SA selection is thought to be a fundamental force forming and maintaining sexual dimorphism [88, 121, 133]. The reason is that intralocus sexual conflict can be mitigated through the independent evolution of the two sexes toward their sex-specific optima [88]. Such resolution of conflict will theoretically result in the establishment of sexual dimorphism [88]. For example, a shared allele

beneficial to males but harmful to females may eventually lead to that allele expressed only in males. In fact, sex-biased expression of shared genes is widely observed [36, 58, 104, 117, 132], and sexual antagonism is a key hypothesis for the evolution of sex-biased gene expression [49]. Sexual dimorphism is common in nature, and it is believed that all sexually dimorphic traits are fundamentally the product of sex-biased gene expression during the development of organisms [171]. SA selection has also been suggested to drive the evolution of the sex chromosomes [27, 97, 137], of gene duplication that leads to sex-biased genes [67, 135], and of genomic imprinting [41, 120], because all can mitigate and/or resolve the conflict between the two sexes.

1.2 Reported evidence for sexual antagonism

In theory, sexual antagonism is inevitable and ubiquitous because, between males and females, the environments of shared variants are so different, and the evolutionary interests never exactly coincide. Since the theoretical foundations of sexual antagonism were developed decades ago [88, 119, 136], accumulated and substantial evidence for such phenomena from nature and from laboratory experiments has been presented.

The reported evidence for the existence of sexual antagonism on shared traits is from a very wide range of species. For example, a study of over 13 year period in a natural population of serins (*Serinus serinus*) [14] showed survival selection on bill morphology differs between the sexes, in that selection on bill depth is

directional in females, but stabilizing in males, presumably resulting from the differences in food utilization between the sexes. In birds, there is also evidence for SA selection on tarsus length [105] and beak colour [128]. Another example of sexual antagonism comes from red deer. A long-term study of red deer (*Cervus elaphus*) in nature [57] showed that fathers with high fitness at reproductive and survival levels tend to have daughters with relatively low fitness. In natural lizard populations, SA selection was reported acting on body size [40] and on survival level [153]. As one of the examples from insects, the fly *Drososophila melanogaster* is a common species used in the laboratory to demonstrate evidence for intralocus conflict between the sexes [30,109,124,127,138]. Many more species from non-human animals and insects with evidence for sexual antagonism have been reviewed in recent literature, such as [18] and [39]. Even in the dioecious plan, sexual antagonism has been demonstrated [17,42,103].

In humans, there has also been evidence of sexual antagonism. In a male homosexual study with Italian sample, Camperio-Ciani et al. [26] reported increased fecundity of maternal female relatives of homosexual probands, by comparing the family trees of homosexuals with heterosexuals. The same finding was replicated later [74], suggesting that shared genetic factors can significantly decrease male reproductive success but increase female fertility.

Other potential evidence of sexual antagonism in humans is from the accumulated reports of sex-specific diseases. Nearly all human diseases are thought to have more or less sex-specific differences in prevalence, course and/or sever-

ity [112]. Classic examples include: Women are much more susceptible to a variety of autoimmune diseases [102, 165]; Cardiovascular disease is overall predominant in men, but has a higher incidence in post-menopausal women compared with men [31]; Alzheimer's disease are more common in women [164], whereas schizophrenia and Parkinson's disease are more common in men [4, 145, 177]. Disease-causing variants are believed to be under purifying selection [16] and will be removed from the gene pool. Thus, sex-specific diseases probably reflect the differential change of fitness between the sexes. However, differences of disease do not necessarily indicate the existence of underlying genetic factors causing differential effects between the sexes, because other differences between the sexes, such as the differences in physiology like sex hormones [12, 114] and in behaviour like exposure to cigarette smoke [11], can also contribute to sex-specific diseases.

In summary, the phenomenon of sexual antagonism in humans has not been as widely studied as in non-human species, and clear evidence of sexual antagonism in humans is still rare. Even in non-human species, so far, the vast majority of reported evidence for sexual antagonism has been based on the observation of negative genetic correlation of fitness on shared traits between the sexes in wild or laboratory populations [18], without knowing the genetics underlying the patterns observed. Since SA selection plays an important role in forming and maintaining sexual dimorphism and in human genomic evolution, and genetic components experiencing SA selection are likely to contribute to human health issues of sex-specific effect, more work is needed to detect SA selection, and

especially to explore the genetic basis underlying SA selection in humans.

1.3 Aims of the study

Only intralocus sexual conflict at the viability level will be considered in this study. Even in the non-human species, less attention has been paid to the sex-differential natural selection arising from differential viability, compared to that from differential fertility [39]. The object of the study is to evaluate whether SA selection at the viability level is detectable in humans, and to explore the genetic basis underlying SA selection if it is detectable.

Single-nucleotide polymorphism (SNP) data from a white European random sample are the main materials employed in this study. Only autosomal SNPs are studied in the thesis. In humans, the two sexes share the same probability of inheriting the same autosomal genotypes from their parents. If a SNP locus experiences SA selection at the viability level, the genotype frequencies at this SNP will differ between the sexes. For example, genotype AA at a shared SNP locus, which decreases male mortality but increases female mortality, will result in the observation of higher proportion for AA in males. Thus the difference in genotype frequencies between the sexes indicates SA selection at the viability level prior to the time of sampling. If SA selection is detected, then the extensive annotation for genes and genomic regions experiencing SA selection will be used to infer the genetic basis underlying SA selection, and the characteristics of variants experiencing SA selection will be explored.

Chapter 2

Genomic data and quality control

Abstract

Autosomal SNP data from a random sample of white Europeans was employed to test whether there were genetic variants undergoing SA selection in humans. The SNP data are from the Wellcome Trust Case Control Consortium (WTCCC), where all individuals were Caucasian and living in Great Britain, and blood samples were collected originally for case-control studies [170]. I used only the data of control cohorts for the genome-wide scan for signals of SA selection. In the raw data, there were six thousand control individuals which were genotyped at ~ 1.8 million SNPs across the whole genome. Given the huge number of SNPs tested in the genome-wide scan, before interrogating each SNP, quality control on individuals and SNPs was performed, because even a small amount of systematic error, when applied across the huge number of SNPs, could still result in a considerable number of false positives which swamp the real signals. After

the quality control, there were $\sim 4,800$ control individuals and ~ 1.2 million SNPs remained for analysis. In this chapter, information on the SNP data and the quality control process are presented.

2.1 SNP data

The genomic data utilized for the analysis of SA selection was from the WTCCC. The WTCCC is a group of 50 research groups across the UK, and aims to identify genetic variants associated with some common diseases [169]. Large-scale samples have been genotyped by the WTCCC, with high-throughput genotyping technologies. The SNP data for control cohorts from WTCCC phase 2 and the SNP data for case cohorts from WTCCC phase 1 were utilized in this study. These SNP data were obtained after application to the WTCCC from <http://www.wtccc.org.uk/>.

2.1.1 Subjects

For both WTCCC control and case cohorts, all individuals were living within England, Scotland or Wales, and the vast majority were self-identified as white Europeans [169]. The two control collections were the 1958 Birth (58C) and UK Blood Services (UKBS) cohorts, each comprising approximately three thousand individuals. The 58C cohort consists of individuals, who were born in the same week in 1958 and provided blood samples when aged 44-45 years. In the UKBS cohort, samples were from blood donors with age range 18-69 years. Both cohorts contain approximately the same number of males and females.

At the time of this study (February 2011), the WTCCC had combined the two control cohorts to form a shared control group for comparison with twenty four common diseases, and more diseases may be included in the future [169]. The two control cohorts are assumed to be samples from the general population. Comparison of SNP allele frequencies between those two control samples revealed few significant differences [170]. As a result, these two control cohorts were also combined in this study.

SNP data for the seven case cohorts from the WTCCC were also employed in this study, to see if the genetically sexual antagonism was associated with diseases. These seven diseases were bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes [170]. The seven case cohorts each comprise approximately two thousand individuals.

2.1.2 The related genotyping technologies

The SNP data from the WTCCC were generated using DNA microarrays. These arrays have millions of probes on a single chip, and scan millions of SNPs per sample. Two kinds of platforms, Illumina 1.2M and Affymetrix V6.0, were employed to genotype each control individual [169]. Genotyping on both platforms starts with a whole-genome amplification step, followed by hybridization between the specific oligonucleotide probe and the target DNA sequence [44, 123]. However, these two platforms differ substantially in several respects:

- Different methods are used to discriminate the two alleles of a SNP. The Affymetrix V6.0 discriminates alleles by direct hybridization of genomic DNA to arrays which contain locus- and allele- specific oligonucleotides [99, 101]. These oligonucleotides are labeled and represent either perfect match or mismatch probes to each SNP. However, the Illumina 1.2M arrays contain only locus-specific probes, and use single-base extension with a labeled base to distinguish the SNPs [68]. Theoretically, the enzymatic fidelity in primer extension used to distinguish the two SNP alleles is extremely high [44, 149], and thus in theory the Illumina technology should show lower background noise.
- The length of probes is different. Illumina arrays use 50-mer oligonucleotides [149], compared to Affymetrix's 25-mers [101].
- The SNP-selection strategies differs. Illumina's probes are based almost entirely on 'tagSNPs', which were those SNPs identified in the HapMap project [76]. In contrast, only about half of the SNP probes on the Affymetrix array are tagSNPs. Although Illumina 1.2M and Affymetrix V6.0 have similar throughput (~ 1 million SNPs per run), they were designed to genotype mostly different SNPs, with only $\sim 18\%$ of the SNPs common to both platforms.

For the case cohorts, a different Affymetrix's platform (Affymetrix 500K), was employed. Affymetrix 500K follows the same procedural steps and features as

Affymetrix V6.0, but the Affymetrix 500K only genotypes about half of the number of SNPs.

2.2 Data quality control

2.2.1 Importance of quality control

The quality control (QC) was performed on the individual level as well as on the SNP level before any analyses were undertaken. The goal of QC was to identify and remove individuals with possible misclassification between males and females or low genotyping quality, and to identify and remove SNPs with potential genotyping artifacts.

The two genotyping platforms which produced the SNP data for control cohorts, the Illumina 1.2M and the Affymetrix V6.0, both claimed accuracy in excess of 99.5% [101, 149]. However, there are at least two reasons why QC for genome-wide SNP data is still important. Firstly, in practice, it is rare to reach genotype call quality as high as the SNP panel manufacturers claim, because the DNA quality of real samples is usually less than the carefully prepared samples used to benchmark the panels [34, 90, 125, 155, 168]. Imperfections involved in any step of the genotyping process, from the sample collection and storage to the operation of SNP genotyping microarrays, can produce genotyping artifacts. Secondly, and perhaps most importantly, in the genome-wide context a small amount of poor quality can have a large impact [168, 170], because even a small amount

of systematic error, when applied across the huge number of SNPs, could still result in considerable number of false positives which would swamp real signals. For instance, the reported accuracy of 99.5% means that ~ 5 thousand (0.5% of ~ 1 million) SNPs have bad quality on each genotyping platform employed in the study, and may be spuriously associated with the trait of interest.

2.2.2 Quality control on individual level

The QC on individuals is intended to identify and remove the individuals that were misclassified or that had low genotyping quality, which may have arisen from sample handling mistakes, contamination, or low DNA sample quality [146].

Along with SNP data, the WTCCC provided the suggested individual exclusion lists, in which there were results for examining the genotyping outcome of each of the individuals against several QC filters. I utilized these exclusion lists, and removed any control or case individuals which failed any of these QC filters. For the control cohorts, two exclusion lists were provided because each of the individuals was genotyped on both Illumina 1.2M and Affymetrix V6.0 platforms, and one may have failed the QC on one platform but passed the QC on the other platform. In this study, if an individual failed QC on any platform, it was excluded.

The following seven filters were used by the WTCCC for the individual QC:

1) High missing data rate. A genotyping result is ‘missing’ when the genotyping platform can not produce a reliable outcome for that genotype, where

the reliability is quantified by the calling algorithm. Individuals with a high rate of missing genotypes are excluded, as this is an indication of low DNA quality [146, 170]. An individual with $>3\%$ missing data across all SNPs was excluded by the WTCCC [170].

2) Outlying heterozygosity. In theory, individuals drawn from a population with random mating should have genotypes in Hardy-Weinberg equilibrium (HWE). The proportion of an individual's genotyped SNPs which are heterozygous is predictable from the Hardy-Weinberg expectation and the allele frequencies at each SNP [168]. Because the probability that an individual has outlying genome-wide heterozygosity is small, outlying genome-wide heterozygosity may indicate contamination [146, 168, 170]. An individual with $>30\%$ or $<23\%$ heterozygosity was excluded by the WTCCC [170].

3) Relatedness. The WTCCC excluded related individuals after running a Hidden Markov Model to identify the proportion of identity by descent (IBD). If two individuals were $>5\%$ identical by descent then the one with the lowest call rate was excluded [169], where call rate is the proportion of SNPs for which a specific allele SNP can be reliably identified.

4) Non-Caucasian ancestry. The WTCCC conducted a principal component analysis on the sample individuals against the HapMap individuals of Caucasian ancestry [170]. Individuals were excluded if they differed from the majority of the collection in terms of ancestry. These excluded individuals were likely of non-Caucasian ancestry.

5) Gender discrepancy. Ideally, a male is expected to have only homozygous genotyping outcomes on the X chromosome due to the sole X chromosome. As for females, both homozygous and heterozygous genotyping outcomes are expected to be observed on the X chromosome. Thus, gender can be inferred from the genotyping outcome of the X chromosome. Individuals with gender discrepancies between the supplier and the inferred gender were excluded by the WTCCC.

6) Outlying channel. The WTCCC excluded individuals if the mean of the intensity values from allele A and allele B from the specified 10,000 SNPs on chromosome 22 was outlying when compared to the sample at large.

7) Genotyping discordance. Prior to being genotyped, the WTCCC genotyped 30 SNPs using the Sequenom equipment at the Wellcome Trust Sanger Institute. After full genotyping, the calls for these 30 SNPs were checked against the Sequenom genotypes. If the concordance was $<90\%$ then the individual was excluded by the WTCCC.

After the QC on individuals, $\sim 2,500$ (84%) and $\sim 2,300$ (77%) control individuals were retained for the 58C and UKBS, respectively. Case individuals failing the QC were also removed from the initial 3,000 individuals in each of the seven case cohorts. The composition for each control and case cohort after QC in this study is listed in Table 2.1.

Table 2.1: Composition of each control and case cohort after quality control

Name of cohort	Males	Females	Total	Sex ratio (males/females)
58C (1958 British Birth Cohort)	1295	1207	2502	1.07/1
UKBS (UK Blood Services)	1151	1163	2314	0.99/1
BD (bipolar disorder)	697	1171	1868	0.60/1
CAD (coronary artery disease)	1527	399	1926	3.83/1
CD (Crohn's disease)	680	1068	1748	0.64/1
HT (hypertension)	775	1177	1952	0.66/1
RA (rheumatoid arthritis)	470	1390	1860	0.34/1
T1D (type 1 diabetes)	998	965	1963	1.03/1
T2D (type 2 diabetes)	1118	806	1924	1.39/1
All Cohorts	8711	9346	18057	0.93/1

2.2.3 Quality control at SNP level

The QC used at the SNP level was designed to identify and remove SNPs that potentially lead to false positives or false negatives. QC processes were applied to SNPs from the control individuals only, as the genome-wide scan was based on the control data. Since the control individuals were genotyped on both of Illumina 1.2M and Affymetrix V6.0 platforms, the same QC filters were executed on each SNP data set from the two platforms, respectively. Below are the details for the SNP QC.

2.2.3.1 Exclusion of the non-autosomal SNPs

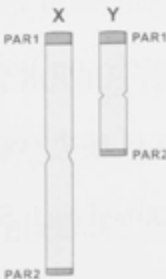


Figure 2.1: Structure of the human male sex chromosomes and the localization of PAR1 and PAR2.

Only autosomal SNPs were scanned in this study for signals for SA selection. The coordinate information for the genotyped SNPs was from the Illumina (www.illumina.com) and Affymetrix

(www.affymetrix.com) companies. Any SNPs that had no available coordinate information or shared the same coordinate with others, were discarded. The SNPs in the pseudoautosomal regions (PARs) were also excluded. PARs are the regions of identical sequence located at the tips of the X and Y chromosomes, with short arms for PAR1 and long arms for PAR2 (Figure 2.1). Their physical lengths are approximately 2.7 Mb for PAR1 and 0.33 Mb for PAR2 [56]. During meiosis, pairing and crossover in men takes place in these two regions like autosomes [24, 60]. However, at the PAR region close to the pseudoautosomal boundary, no such recombination occurs [56]. In this case, a man will always pass the alleles in the PAR of X chromosome to his daughters and the alleles in the PAR of Y chromosome to his sons, and consequently no concordant genotype distributions between the sexes is expected. The SNP data in my study showed such discordance between males and females in the PAR (Figure 2.2). Because it was not practicable to define the SNPs neighbouring the pseudoautosomal boundary, all the 800 genotyped SNPs in PARs were discarded.

2.2.3.2 Common QC filters from the WTCCC

The WTCCC genotyped the two control cohorts separately, on both Illumina 1.2M and Affymetrix V6.0 platforms. For the SNP data of each control cohort generated on each genotyping platform, the WTCCC examined each SNP against

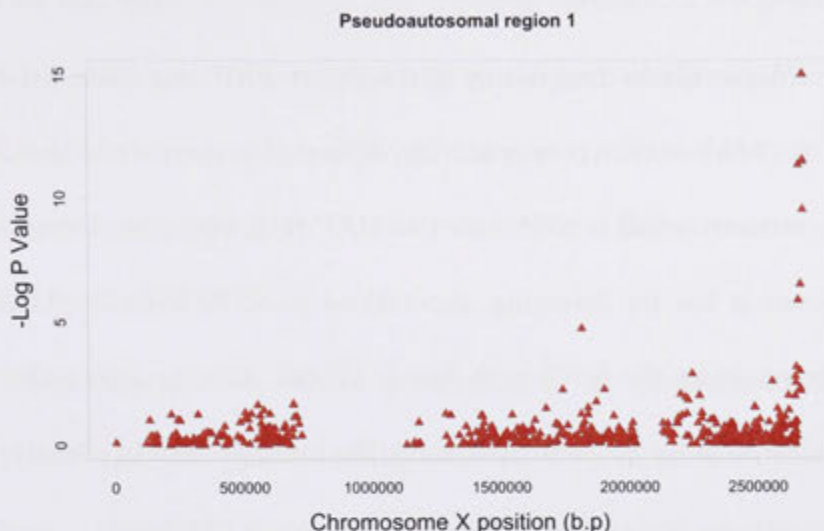


Figure 2.2: The significant discordance in the genotype distribution between the sexes at the SNPs close to the pseudoautosomal boundary.

X-axis is the genomic location, which here is chrX:1-2,709,520 for the whole pseudoautosomal region 1 according to the UCSC genome group, in NCBI-36 coordinates; Y-axis is $-\log(\text{P value})$. Red triangles represent single SNPs typed in the study.

five QC filters and provided the results. According to these results, SNPs failing any of the five filters in any control cohort were discarded from this study. The majority of these filters were also common in other genome-wide studies, and were designed mainly against those genotyping artifacts which were caused by human errors such as sample contamination and genotyping misoperation [168]. The implementation of these WTCCC's filters excluded $\sim 20.4\%$ of the autosomal SNPs genotyped on Illumina and Affymetrix platforms respectively. Below are the details for these five QC filters.

1) High missing data rate. The WTCCC claimed that the best indicator of a SNP being difficult to call was the amount of missing data in its calls [170]. A SNP was discarded, when the proportion of individuals for which the SNP can

not be reliably identified was greater than 0.02.

2) Low minor allele frequency (MAF). A SNP was excluded by the WTCCC if the MAF $< 1\%$. It is relatively difficult for the current genotyping platforms to accurately call a SNP with low MAF [110, 168, 179]. Moreover, the statistical power is low for detecting associations to SNPs with low MAF, and consequently, retaining the SNPs with low MAF can directly reduce the power to detect signals in other SNPs by increasing the multiple testing penalty [168].

3) Departure from Hardy-Weinburg equilibrium. Testing for departures from the assumption of Hardy-Weinberg equilibrium (HWE) has been widely used as a criterion for identifying SNPs with genotyping errors in genetic case-control studies [72, 92, 142, 156, 168, 178]. Gross genotyping errors may arise among the genotypes called using automated and unsupervised procedures including the genotyping technologies used to produce the WTCCC data. However, a SNP under sex-differential natural selection can also cause departures from HWE, unless complete neutralization between the two sexes occurs. In this study, the SNPs failing the WTCCC's QC filter ($P < 1E-20$ for HWE testing) were excluded to produce high-quality data, but any SNPs solely failing this filter were examined carefully to ensure no genuine signal had been missed (for results see Chapter 3.2.2.1).

4) Information measure. This is a measure of the amount of information about the allele frequency present in the genotype call probabilities, relative to the expected amount when the genotypes were known with certainty. It is cal-

culated by one minus the ratio of the expected variance of the genotype (given the genotype calls) and the variance you would get for a Hardy-Weinberg SNP if only the allele frequency is known [98]. As genotype uncertainty increases, the information measure goes down. A SNP was excluded if this measure < 0.975 , as suggested by the WTCCC.

5) Plate association. Gross genotyping errors may happen to the SNPs genotyped on the same plate, leading to genotyping errors associated with specific plates. The WTCCC conducted a test for the plate association with the genotyping outcome, and a SNP was excluded if $P < 1E-5$.

2.2.3.3 QC against SNPs in segmental genome duplications

A large amount of segmental duplications (duplicons) are distributed across the whole human genome [29, 162]. Duplicons, defined as stretches of DNA larger than 1 kb and over 90% similar between copies, comprise $\sim 5\%$ of the human genome [9, 29, 162]. This proportion will be increased, if the size restriction is reduced.

SNPs can be located in duplicons. Due to reduced selective pressure, mutations increase their frequencies more easily in duplicon regions, resulting in relatively high density of SNP [35]. SNP databases report that SNPs are over-represented by a factor of ~ 2 in duplicon regions [29, 50]. In fact, this figure is underestimated because SNP discovery efforts have discarded predicted variants from regions where a duplicon is suspected [141, 160].

The probe hybridization process of current genotyping microarrays can not distinguish the duplicons, with probes matching and hybridizing to all of the duplicons, and thus return the combination of genotyping multiple genetic locations. Due to the difference in sex chromosomes, duplicon SNPs between autosomes and sex chromosomes can lead to differential genotyping results between the sexes, causing false positives. Consider the following hypothetical example, where there is a segmental duplication between the autosome and Y chromosome. A SNP with alleles A and a is located in the autosomal copy of the duplicated region, but only allele a is in the corresponding locus on the Y chromosome. Males with autosomal genotype AA will be mistakenly assigned genotype Aa due to the allele a detected in Y chromosome, while females with the same genotype AA in autosome will be still assigned genotype AA due to the lack of Y chromosome.

To detect those SNPs which were potentially affected by duplicons between autosomes and sex-chromosomes, two measures were applied against short and long duplicons respectively. Firstly, the specific oligonucleotide probe sequence of each autosomal SNP was aligned against the X and Y chromosomal sequences, using BLAT [80]. This identified autosomal SNPs where hybridization between the probe and the sex-chromosomal DNA sequence could possibly occur. Probe sequences were from the Illumina and Affymetrix companies, downloaded via the WTCCC official websites (www.wtccc.org.uk). The sex-chromosomal sequences (GRCh37 version) were downloaded from the UCSC Genome Bioinformatics group, at <http://hgdownload.cse.ucsc.edu/downloads.html#human>.

For increased sensitivity of BLAT search, I used the parameters, `-tileSize=6`, `-stepSize=2` and `-repMatch=1,000,000`, for the probe sequences (50-mer oligos) of Illumina 1.2M array. The parameter `tileSize` defines the size of a match ('tile') that triggers an alignment, `stepSize` is the spacing between tiles, and setting `repMatch` to a large value reduces the chance of a tile being marked as over-used and ignored. The formula to calculate the shortest size that can guarantee a match is: $2 * \text{stepSize} + \text{tileSize} - 1$. So these parameter setting could guarantee an exact match as short as 9bp. As for the short probe sequence (25-mers) of Affymetrix V6.0, the same parameterization was used, except `-stepSize=1`, which could guarantee an exact match of 7bp. Incomplete hybridization between the probe and sex-chromosomal sequence, which looks like an upside-down 'U' (Figure 2.3), was assumed to be able to affect the genotyping. The upside-down 'U' is represented by continuous gap characters in the BLAT results. A hit with match of $\geq 80\%$ of the length of probe sequence, i.e. 40bp for Illumina 1.2M array and 20bp for Affymetrix v6.0 array, and with no more than one insertion of continuous gaps of ≤ 1000 bp, would lead to the corresponding SNP being discarded.

The second measure was against the SNPs located in long duplicons. This step

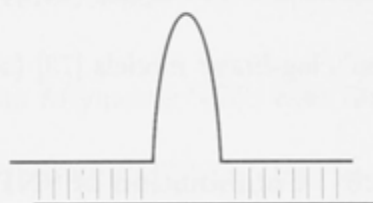


Figure 2.3: The incomplete hybridization between the probe and the genomic sequences. Any one of the two bold lines represents the probe sequence, and the other represents the target genomic sequence. The vertical thin lines represent the bonding between the probe and the genomic sequences, with the solid for match and the broken for mismatch.

complements the step described above, as

the BLAT search may miss some duplicons due to the imperfection of the genomic sequencing. The definition data for long duplicons, which were defined as being >1 kb with $>90\%$ similarity between copies, were from Bailey et al. [10] with a update for NCBI36 genome version. SNPs located in the autosomal regions which had long duplicons at X or Y chromosome were discarded.

2.2.3.4 QC for the combination of control samples

As discussed in Section 2.1.1, the 58C and UKBS cohorts were combined to improve statistical power. Therefore, SNPs with $P < 1E-7$ for the difference in genotype distributions between the two control groups were excluded. The discrepancy might be an indicator for the genotyping errors occurring in any one or in both of the two cohorts. The P values for comparison of genotype distributions between the two control groups were generated with Huttley and Wilson's log-linear models [73] (see Chapter 3.1.1 for description).

2.2.3.5 Combination of SNP data from the two genotyping platforms

As mentioned in Section 2.1.2, each individual from control cohorts was genotyped on both of the two platforms, the Illumina 1.2M and the Affymetrix V6.0. There were initially ~ 1.15 million SNPs on Illumina 1.2M and ~ 0.93 million on Affymetrix V6.0, with approximately 0.3 million genotyped on both platforms. For those SNPs genotyped on both platforms, only the outcome from Illumina's were employed for analysis of SA selection. The Illumina platform was chosen for

three reasons:

1. The method that Illumina employs to discriminate the two alleles of a SNP may show a lower background noise (see Section 2.1.2).
2. Illumina's longer probe sequences can increase hybridization specificity. This reason may be important if the interrogated SNP is located in short duplicons. For example, assuming an autosomal SNP, which is targeted for genotyping, is located in a region completely duplicated in a sex chromosome, and the length of these duplicons is 25-bp. The Affymetrix's probes of 25-mer will hybridize to both duplicons with 100% matching, and thus can not distinguish the duplicons. However, the sex-chromosomal duplicon may not disturb the Illumina platform, because for the 50-bp of Illumina's probes, only 50% matching to the sex-chromosomal duplicon can lead to the stop of genotyping process on the sex chromosome.
3. Due to shorter length of probes, many more Affymetrix SNPs were filtered out by the QC filter of BLAT search in this study (22.5% of all autosomal SNPs excluded from Affymetrix's platform vs. 1.3% of those from Illumina's. See Tables 2.2).

2.3 Summary

QC is a key step for genomic data analysis, serving the critical function of reducing the number of false positives and false negatives in the subsequent analyses. Strict

and comprehensive QC on both individual and SNP levels has been performed to ensure the high quality of data utilized in the analysis of SA selection. Firstly, I employed all the QC filters developed by the WTCCC. These QC measures are used frequently as standards in genome-wide association studies [90, 146, 155, 168], aiming to exclude the individuals which accompany inadvertent sample duplication, relation or contamination, or low DNA quality, and the SNPs involved with potential genotyping errors. The QC on the individual level was finished after this process. Some of the original SNPs were also filtered out. Secondly, specific and novel QC steps against sex-chromosomal duplicon SNPs were utilized. These QC measures were developed to detect the SNPs located in duplicons between the autosomal and sexual chromosomes. The duplicon SNP can not be genotyped correctly by the current technologies such as Illumina 1.2M and Affymetrix V6.0 employed by the WTCCC to produce the SNP data in this study. The QC against duplicon SNPs was critical to this study, because duplicon SNPs between the autosome and the sex chromosome can directly lead to false positive signals for SA selection. These novel QC measures can also be applied to other genome-wide studies especially where the two compared groups have marked differences in the ratio of the two sexes. Lastly, to ensure the constancy of genotyping outcomes between the two combined control groups, SNPs showing great difference ($P < 1E-7$) in genotype distributions between the two control groups were excluded.

After QC and the combination of SNPs from the two genotyping platforms, 1,222,210 SNPs were retained for analysis of SA selection. Table 2.2 shows the

proportion of SNPs excluded by each QC filter in this study, as well as the overall proportion, on Illumina 1.2M and Affymetrix V6.0 platforms, respectively.

The validity of the employed QC measures was demonstrated by the high enrichment of probable false positives among the discarded SNPs. Table 2.3 shows the proportion ratios of positives between filtered and unfiltered autosomal SNPs for each QC filter, where positive was defined as $P < 1\text{E-}5$ for sex difference in genotype distribution, the threshold used for the genome-wide scan after QC. The SNPs failing each QC filter contained a much higher proportion of probable false positives compared to the background proportion (the proportion for original SNPs before QC), with the ratio up to 242. The only exception was for the filter against SNPs with low minor allele frequency on the Affymetrix platform, probably due to the low statistical power.

Table 2.2: Number of SNPs failing each QC filter

Filter	Genotyping platform	
	Illumina 1.2M*	Affymetrix V6.0*
High missing data rate	26,836	33,256
Low minor allele frequency	185,460	140,067
Hardy-Weinburg test	19,472	56,127
Information measure	157,953	133,748
Plate association test	100,321	32,692
Differ between 58C and UKBS	16,614	642
Short duplicon (BLAT search)	14,087	199,777
Long duplicon	1,434	742
Failing any of the above filters	235,646	338,281

*The numbers of original autosomal SNPs before QC were 1,115,038 and 887,688 from Illumina 1.2M and Affymetrix V6.0, respectively.

Table 2.3: Proportion ratios of positives between filtered and unfiltered SNPs*

Filter	Genotyping platform	
	Illumina 1.2M#	Affymetrix V6.0#
High missing data rate	29.8	14.7
Low minor allele frequency	2.6	0.6
Hardy-Weinburg test	30.0	6.4
Information measure	5.9	3.1
Plate association test	10.0	15.3
Differ between 58C and UKBS	32.4	155.8
Short duplicon (BLAT search)	28.6	3.0
Long duplicon	148.4	242.6
Failing any of the above filters	4.6	2.5

*Positives were defined as $P < 1\text{E-}5$ for sex difference in genotype distribution, the threshold used for genome-wide scan after QC.

#The positive proportions in the original autosomal SNPs before QC were 4.7E-5 and 1.1E-5 on Illumina 1.2M and Affymetrix V6.0 platforms, respectively.

Chapter 3

Evidence for sexually antagonistic selection in humans

Abstract

Sexual antagonism has been barely investigated in humans, let alone the localization of genomic loci undergoing SA natural selection. In this study, significant differences in genotype distribution between males and females from random samples were used as indicators for sexual antagonism, or sex-differential viability, since the two sexes share the same probability of inheriting the same autosomal genotypes from their parents. This chapter presents the results for the comparison of genotype distribution between the sexes at each of ~ 1.2 million autosomal SNPs across the whole genome, with control samples ($\sim 4,800$ of sample size, with approximately equal numbers of males and females) from the WTCCC passing the quality control process. The results demonstrate evidence for ongoing SA

selection in humans, with eight genomic regions exhibiting a signal with P value $< 1\text{E-}5$. Candidate genes associated with sexual antagonism in those eight regions are identified. The reliability of the positive signals revealed in this study is demonstrated.

3.1 Statistical methods

Differences in genotype distributions between males and females were used as indicators for SA selection. In the genome-wide scan for the signals for SA selection, ~ 1.2 million statistical tests were performed simultaneously for difference in genotype distributions between the sexes. Huttley and Wilson's log-linear models [73] were used to measure such sex difference. To solve the issue of multiple testing, the ~ 1.2 million test results were adjusted in two ways, with calculation of the false discovery rate (the 'q value' method) and the effective number of independent tests (the 'simpleM' method). These statistical methods are introduced below.

3.1.1 Huttley and Wilson's log-linear models

Huttley and Wilson's log-linear models [73] were employed to measure the difference in genotype distributions between the sexes at each of the SNPs in the genome-wide scan. These models compare groups based on the deviations from Hardy-Weinberg equilibrium (HWE) of each group.

For illustration, consider a locus where there are two alleles A and a , with p for

the frequency of allele A and q for the frequency of allele a , and the genotypes are AA , Aa , and aa . The expected genotypic frequencies under HWE are $P_{AA} = p^2$ for AA , $P_{Aa} = 2pq$ for Aa , and $P_{aa} = q^2$ for aa , with $p^2 + 2pq + q^2 = 1$. Thus, under HWE, $4P_{AA}P_{aa}/P_{Aa}^2 = 1$. Let coefficient $M_{AA} = 4P_{AA}P_{aa}/P_{Aa}^2$, and the deviation of M_{AA} from 1 is a measure of the departure of observed genotypic distribution from the expected HWE.

When testing two groups for concordant equilibria, according to Huttley and Wilson, more terms can be added to the log-linear model, written as:

$$\ln(P_{AA})_i = \ln M + \ln \tau_i + 2\ln M_A + \ln M_{AA} + 2\ln(\tau M_A)_i + \ln(\tau M_{AA})_i$$

$$\ln(P_{Aa})_i = \ln 2 + \ln M + \ln \tau_i + \ln M_A + \ln(\tau M_A)_i$$

$$\ln(P_{aa})_i = \ln M + \ln \tau_i$$

where $i = 1, 2$ for the two groups respectively. The parameter M is the differences between observed numbers of genotypes (AA , Aa , and aa) and the mean value of these numbers in the total samples; τ represents the possible unequal sizes of samples between two compared groups; M_A represents the frequency of allele A compared with allele a in the total sample; M_{AA} represents the Hardy-Weinberg (HW) disequilibrium coefficient in the total sample; τM_A represents the difference in allelic frequencies between the two groups; and τM_{AA} , the difference in Hardy-Weinberg (HW) disequilibria between the two groups. The deviance for genotypic concordance between the two compared groups is the sum of values of τM_A and

τM_{AA} , which is the measure for SA selection in this study. The calculation was done with an R package, ‘hwde’, comprising these models and developed by J.H. Maindonald at <http://cran.r-project.org/web/packages/hwde/index.html>.

3.1.2 Adjustments for multiple testing

In statistical significance testing, P value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true [143]. In the case of comparison of genotypic distributions between the sexes at a SNP in this study, the null hypothesis is that there is no difference in genotypic distribution between the sexes. However, random sampling from males and females can still provide discordance under the condition of true null hypothesis. The distribution of P values generated by statistical testing with random samples under null hypothesis is continuous and uniform on the interval $(0, 1)$ [116]. So P values can be used to measure the strength of evidence against the null hypothesis. For example, $P = 0.01$ simply means that the observed difference can be attributed to chance 1% of the time when the null is true, and a very small P value provides strong evidence that null hypothesis is unlikely to be true. Likewise, choosing 0.05 of P value as the threshold of significance means there is a 5% probability of obtaining a significant result when the null hypothesis is true. A more stringent threshold reduces the probability that significant results are obtained simply by chance when the null hypothesis is true.

Multiple testing problem occurs when a set of statistical inferences is considered simultaneously. As discussed above, the distribution of P values generated by statistical testing under true null hypothesis is uniform on the interval $(0, 1)$. Hence, under the null hypothesis, five positive results are expected to appear simply due to chance, when one hundred testes are performed simultaneously with $P = 0.05$ selected as the threshold. Given the ~ 1.2 million of SNPs tested in the genome-wide scan in this study, a huge number of positives can appear simply by chance.

For this reason, strict significance thresholds have been recommend to control the experiment-wise error rate in multiple testing. One of the traditional and popular methods is the Bonferroni correction [19], which controls the overall false positive rate (e.g., $\alpha = 0.05$). In a study of n independent tests, the Bonferroni-corrected α' for each individual test is $\alpha' = \alpha/n$. So for the 1.2 million testes, to reach 0.05 of overall false positive rate, the Bonferroni-corrected $\alpha' = 0.05/1.2E6$, which is $5E-8$. However, the Bonferroni correction method is overly conservative in genome-wide studies [107, 134, 167], where hundreds of thousands of testes are not completely independent due to linkage disequilibrium (LD) relationships, leading to a huge number of false negatives. To resolve this statistical issue, the following methods for correcting the multiple test results were employed in this study.

3.1.2.1 Q value for false discovery rate

The first method employed in this study to resolve the multiple testing issue is to calculate the q value for false discovery rate (FDR). FDR is the expected proportion of false positive findings among all the rejected null hypotheses [151]. The q value is defined to be the FDR analogue of the P value, and is the minimum FDR at which the test may be called significant [151]. In the study, the q value for each test was calculated according to Storey et al [150, 151], after obtaining all P values for the sex differences in genotype distributions in the genome-wide scan by Huttley and Wilson's log-linear models described above. The principal algorithm for this mathematical method is to calculate the proportion of false positives by the comparison of distributions between the observed P values and the P values from null hypothesis (Figure 3.1). Storey et al. claimed that this method also works well with multiple testing of weak dependence [151]. It is one of the most popular methods for multiple testing adjustments, with more than four thousand citations according to Scopus literature database [93] (up to February 2011). The calculation was performed with the R package, 'qvalue', from <http://cran.r-project.org/web/packages/qvalue/index.html>.

3.1.2.2 The simpleM correction

The simpleM correction [64] is a recently published method, which is based on the estimation of the number of independent tests from genome-wide SNP data. In this method, the genotypes of neighboring SNPs of all individuals are employed to

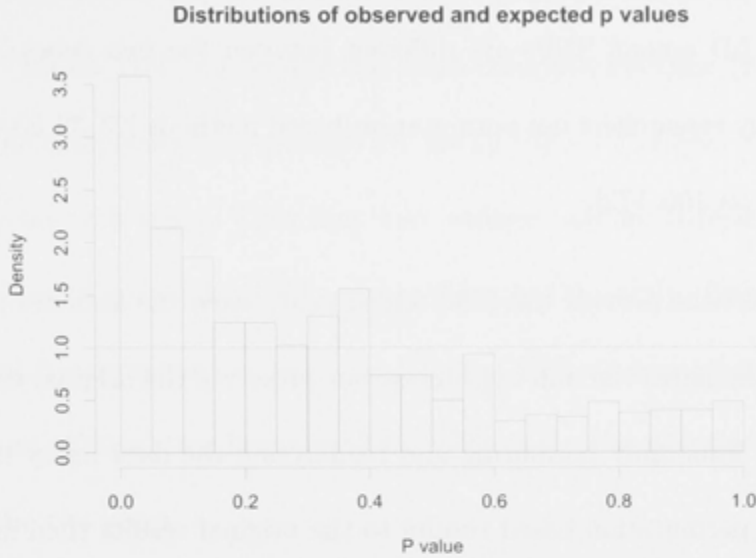


Figure 3.1: Calculating the false discovery rate for multiple testing correction. The bars represent the distribution of observed P values from a multiply tested hypothesis; The dashed line is the density histogram that is expected under the null hypothesis. The proportion below the dashed line can be considered as false positives when they are called ‘significant’.

infer the linkage disequilibrium (LD) correlation structure of SNPs, and therefore the effective number of independent tests can be estimated. Standard Bonferroni correction then can be used to adjust the multiple testing. The program of simpleM was downloaded at <http://simplem.sourceforge.net/>.

3.1.2.3 The permutation test

The two above methods for multiple test adjustments were used in the genome-wide scan where ~ 1.2 million tests were performed simultaneously. In the case of adjustment of the P values from SNPs located in a single gene, permutation test were employed for more accurate adjustment. For instance, to choose the gene with stronger signal for SA selection from two genes, we cannot simply compare

the two smallest P values from SNPs in the two genes, because the number of SNPs and the LD among SNPs are different between the two genes. To solve such issue, many researchers use permutation-based methods [23,32,55,61,79,87,91,94,95,111,139,166,172].

Permutation tests provide the valid adjustment, when the data are permuted in a way that simulates the null hypothesis but preserves the original correlation structure [32]. Randomly permuting and reanalyzing the data many times and comparing the permutation-based results to the original results then allow estimation of the probability of achieving the originally observed P value, given the correlation between tests.

In this study, in each permutation shuffle for a gene, the SNP genotypes of each individual were kept the same, while the sexes of each individual were shuffled. Such shuffle can maintain the haplotype of each individual and thus the LD among SNPs, and simulate the null-model of no sex-differential natural selection. For each permuted sample, the difference in genotype distribution between the sexes at each SNP was tested, and the smallest P value was recorded. I performed 10,000 permutations, as Churchill and Doerge suggested [32], and had 10,000 smallest P values for a gene. The 10,000 smallest P values were then arranged in ascending order, and the percentile where the observed smallest P value from the original WTCCC sample was located would be presented as the adjusted P value for that gene. The permutation-adjusted P values from different genes could be compared to choose the gene with the strongest signal for SA selection.

Permutation tests are considered the gold standard for multiple testing correction [37, 63, 69, 115], but require too much computation time (years estimated to finish the computation on a single PC for all the ~ 1.2 million SNPs employed in the genome-wide scan). Thus they were utilized only for adjustment of signals from a subset of genes identified as interesting in this study. Likewise, to make it practical for computation time, here I used Fisher's exact test, rather than Huttley and Wilson's log-linear models, to compare the genotype distributions between the two sexes.

3.2 Results

3.2.1 Examination of reported variants with sex differences in genotype distribution

So far, to my knowledge, only two genetic variants had been reported to show significant difference in genotype distributions between the sexes in general human populations. I consider these two variants before presenting the results from the genome-wide scan for signals of SA selection in this study. The data here are the same as for the genome-wide scan. The first variant is SNP rs144848 in BRCA2 (breast cancer 2, early onset; Ensembl gene ID: ENSG00000139618) on chromosome 13. At this SNP, Healey et al. [71] originally reported a significant difference between genotype distributions of the unselected newborn males and females in the UK ($P = 0.0015$, $\sim 1,150$ of sample size for each sex), with an

Table 3.1: Genotype distributions of rs1801406 in BRCA2

	Observed number of genotype			Allele frequency		Deviation from HWE
	AA (%)	AG (%)	GG (%)	A	G	
Males	1136 (46.6%)	1100 (45.1%)	201 (8.2%)	69.2%	30.8%	$P = 0.004$
Females	1128 (47.7%)	986 (41.7%)	253 (10.7%)	68.5%	31.5%	$P = 0.095$
Combined	2264 (47.1%)	2086 (43.4%)	454 (9.5%)	68.8%	31.2%	$P = 0.420$
Sex difference in genotype distribution: $P = 0.004$						

excess of heterozygotes in females but a deficit in males, compared to the expected HWE respectively. Later, Teare et al. [154] repeatedly observed the significant heterozygote excess in females from control populations, but found no evidence of genotype specific selection in males, and no statistical evidence that the male and female viabilities were genotypically different from each other. This SNP was genotyped by the WTCCC on the Illumina platform and passed all QC filters in this study. However, there was no significant signal ($P = 0.30$) when I compared the genotype distributions between the two sexes from WTCCC’s control samples. The P values of Fisher’s exact test of HWE were 1.00 for females, 0.66 for males, and 0.72 for the combined sexes, all showing no violation of HWE.

Healey et al. [71] also reported that the homozygous genotype (C/C) at rs144848 in BRCA2 was associated with a 30% increase in breast cancer risk, in a case-control study of Caucasian European samples. This finding was later confirmed by another independent study [147], but has not been replicated in several other studies [22,38,59,65]. Thus, it was suggested that the allele at this SNP locus might be a low-penetrance risk factor for developing breast cancer [131], or

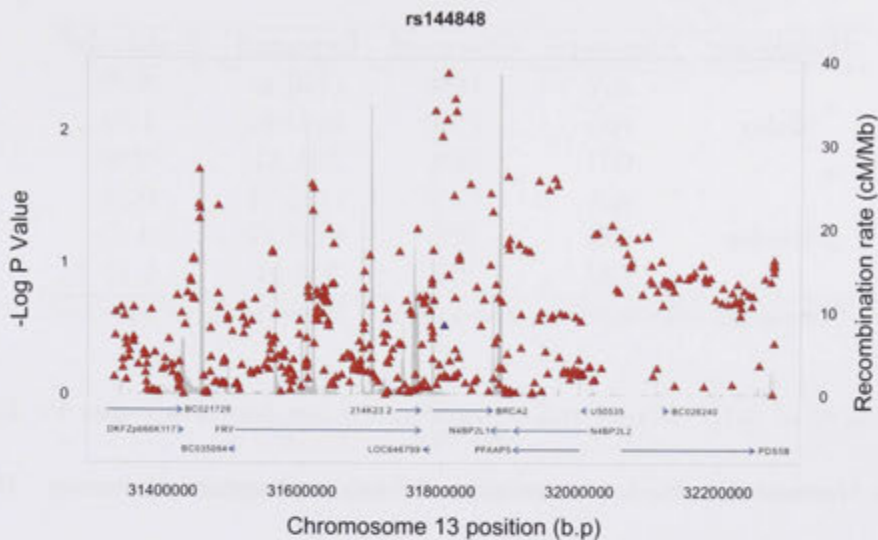


Figure 3.2: The distribution of signals for SA selection across a 1 Mb genomic region centered on BRCA2.

X-axis is genomic location, in NCBI-36 coordinates; y-axis, left, is $-\log(P \text{ values})$, and right, is recombination rate. Red triangles represent single SNPs employed in the study. Rs144848, for which a significant association with sex-differential prenatal viability is reported [71], is indicated by a blue triangle. The SNP showing the strongest signal is rs1801406 ($P = 0.004$). Other SNPs with $P < 0.01$ are rs2320236, rs9534174, rs9567578 and rs4942448. Grey bars represent recombination rates. In lower panel, blue arrows are the genes in this region, and gene annotations were taken from the UCSC Genome Browser.

that this SNP is in LD with the causative variant rather than being the causative variant itself [71]. These explanations may also apply to the failure to observe sex difference in genotype distributions in this study, if the degree of association with breast cancer is directly proportional to the degree of sex specific selection.

Although rs144848 may not be strongly associated with breast cancer, BRCA2 is a tumour-suppressor gene [20, 28, 144], and deleterious mutations in BRCA2 are a major cause of hereditary breast and/or ovarian cancer [48, 52, 77, 83, 106, 152, 176]. A wide range of types of mutations, including SNPs, insertion-deletion and larger genomic rearrangements, has been reported for BRCA2 (reviewed by

Table 3.2: Observed and expected genotype frequencies at rs1801406 in BRCA2

Population	Genotype	Observed	Expected	Residuals [#]
Males	AA	1136	1166.43	-0.89
	AG	1100	1039.14	1.89
	GG	201	231.43	-2.00
Females	AA	1128	1110.11	0.54
	AG	986	1021.77	-1.12
	GG	253	235.11	1.17

[#]Residual calculated as $(Observed - Expected)/\sqrt{expected}$

Fackenthal et al. [51]). Breast and ovarian cancer are mostly sex-specific diseases, that can theoretically lead to genotypic differences between the sexes. BRCA2 is thus a good candidate gene for SA selection, and I decided to examine all the WTCCC’s SNPs located in this gene.

Forty SNPs located in BRCA2 (~84 Kb length) passed the QC. Among these SNPs, rs1801406 shows the strongest signal for sex difference in genotype distribution with $P = 0.004$ (Table3.1). Permutation tests with the 40 SNPs produced adjusted $P = 0.075$. The differences between the sexes at BRCA2 are also the largest across the region of 1 Mb centered on BRCA2 (Figure3.2). These results suggest that sex-differential natural selection is acting on BRCA2, and that rs144848 is not the causative variant of the previously reported positive association with sex-differential prenatal viability. Despite great differences in genotypic distributions at rs1801406, there is no significant sex difference in the allelic distribution ($P = 0.456$), and departures from HWE were observed in each sex ($P = 0.004$ for males and $P = 0.095$ for females) with relative concordance with HWE expectations ($P = 0.420$) for the combined sexes, suggesting the possible balanced allelic frequencies maintained by sexually antagonistic natural selection.

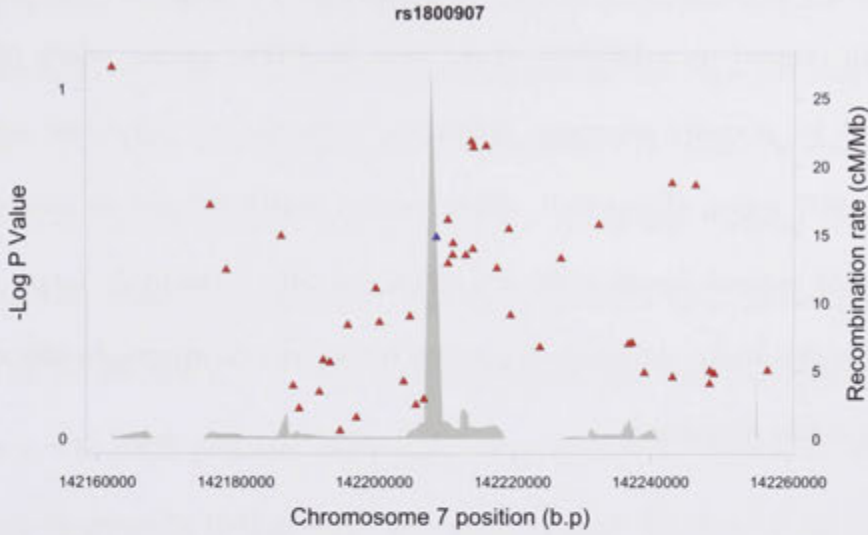


Figure 3.3: Signals for sex differences of genotype distribution in the 100 Kb genomic region with central rs1800907 in blue. This SNP reportedly showed significant signal for SA selection [159].

The comparison between observed and expected genotype frequencies (Table 3.2) also shows the opposite departures between the sexes from the HWE on all three genotypes, indicating the existence of sexually antagonistic natural selection.

The other genetic variant, for which a sex difference in genotype distribution has been reported in general human populations, is rs1800907 on chromosome 7. At this SNP, Tripputi et al. [159] genotyped 200 male and 300 female healthy Caucasian Europeans in Italy, and found significant discordance of genotype distribution between the sexes ($P = 0.012$), with relative excess of heterozygotes in males. This SNP was also genotyped by the WTCCC on the Illumina platform and passed all QC filters in this study. However, no significant signal ($P = 0.26$) was observed when the genotype distributions between the sexes were compared in the control individuals, with P values for the exact test of HWE greater than

0.23 in males, females and the combined sexes. In the genomic region of 100 Kb length centred on rs1800907, there were 44 SNPs, among which all have $P > 0.085$ for sexually genotypic difference, indicating no significant signal in the rs1800907 region (Figure3.3). Since neither rs1800907 nor its surrounding SNPs in LD showed signature for sex difference, plus a relatively larger cohort was used in this study, the previous positive report may be geographically specific or simply a false discovery.

3.2.2 Evidence for SA selection revealed by the genome-wide scan

3.2.2.1 The positive results from the genome-wide scan

In the genome-wide scan, the genotype distribution between the sexes was compared at each of the 1,222,210 million autosomal SNPs across the whole genome. As described in Chapter 2, the individuals were white Europeans from the control cohorts of the WTCCC, with sample size of 4,816 and approximately the same number in males and females, and the SNP data had passed the process of quality control.

Comparison of the observed statistics from the genome-wide scan and the expected values revealed true associations. If there were no sex differences in genotype distributions at any SNP, the ~ 1.2 million test statistics from the genome-wide scan would fit the chi-squared distribution. The comparison between the observed and expected statistics was performed and showed by a quantile-quantile

plot (Q-Q plot) in Figure 3.4. In a Q-Q plot, observed test statistics are ranked in the order from smallest to largest, and plotted against the expected values in the same order that have been sampled from a distribution of known theoretical form (in this case, the chi-squared distribution) [6]. If the distributions of observed and expected statistics are similar, then the points in a Q-Q plot will approximately lie on the line $y = x$. The true associations can lead to deviations from the line $y = x$. An R package, ‘snpMatrix’, developed by Clayton et al. [33], were employed to generate Q-Q plots in this study. Figure 3.4 showed the Q-Q plot from the genome-wide scan in this study, where most of the observed statistics were located in the 95% concentration band, except outliers in the extreme tail. The latter suggested that sexual antagonism, which could cause sex differences in genotype distributions, exists in humans. The distribution of test statistics from all SNPs across the whole genome is shown in Figure 3.5.

The genome-wide scan revealed 19 SNPs with $P < 1\text{E-}5$. The cutoff level of $P < 1\text{E-}5$ was chosen to identify the SNPs with the strongest signals, because the outliers from the 95% concentration band in the Q-Q plot were approximately greater than 22 in Figure 3.4 (corresponding to $P < 1.7\text{E-}5$ with 2 d.f.). Among these 19 SNPs, eight independent signals were identified by the threshold of pairwise LD statistic $r^2 < 0.2$, or the distance greater than 250 Kb between the SNPs when the corresponding r^2 values were absent. The r^2 data were from the HapMap project [76], ranging from 0 to 1, with 1 indicating the complete LD. Table 3.3 shows the test results and the genotype details for these eight

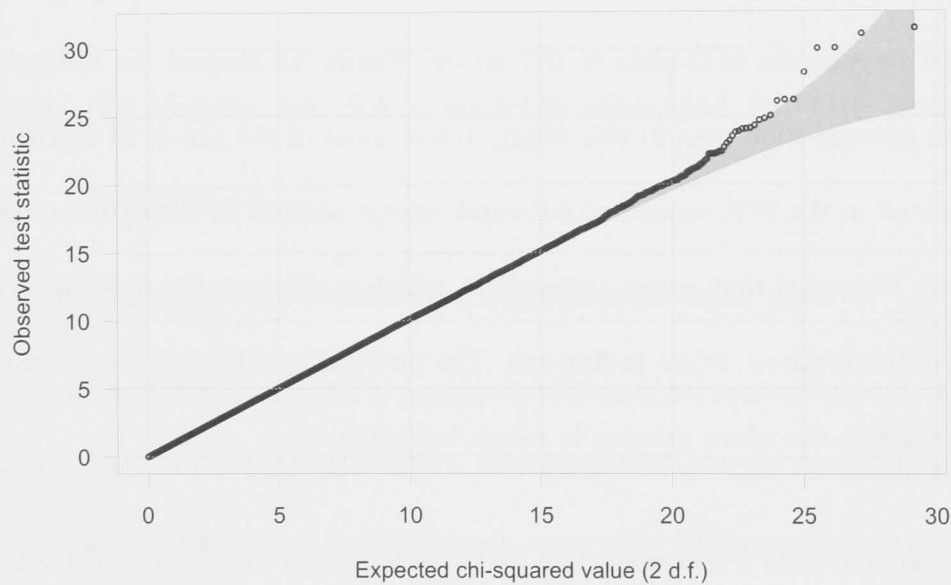


Figure 3.4: Quantile-quantile plot for the test statistics from the genome-wide scan. SNPs failed by the quality control filters were already excluded. The shaded region is the 95% concentration band.

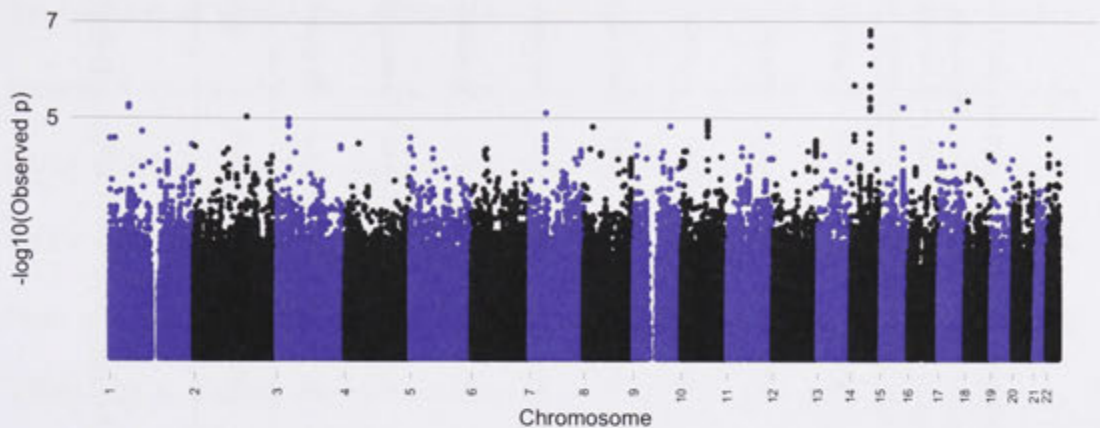


Figure 3.5: Genome-wide picture of the signals for SA selection. The two horizontal lines represent P values of $1E-5$ and $1E-7$ respectively. SNPs failed by the quality control filters were already excluded.

SNPs. Figure 3.6 shows genomic regions of 800 Kb centered on these eight SNPs, including surrounding SNPs and genes within these regions.

The smallest P value from the genome-wide scan was $1.52E-7$ at rs11159094 on chromosome 14. The corresponding q value was 0.096. The effective independent number of tests in the genome-wide scan was 528,486, estimated by the simpleM method. Given the number of independent tests, the adjusted value of the smallest P is 0.080, and the expected number of signals $< 1E-5$ obtained by chance was five but eight were actually obtained from the genome-wide scan.

That the smallest adjusted P and q values were greater than 0.05 might reflect the limited power of the samples used in the study. The sample size of $\sim 4,800$ in the study had only 47% statistical power to detect rs11159094, the SNP showing the strongest signal, with 0.05 of adjusted P value as threshold (corresponding to a $P = 4E-8$ before adjustment by Bonferroni method). To obtain 80% or 95% power, the sample size would need to be increased to 6,500 or 8,100, respectively.

Table 3.3: Eight independent SNPs with $P < 1\text{E-}5$ for the comparison of genotype distributions between the sexes revealed by genome-wide scan

SNP	Chr:Position	P value	q value	HWE P	Alleles (Aa)	Minor allele	AA PR (95% CI)	Aa PR (95% CI)	aa PR (95% CI)	Obs AA	Obs Aa	Obs aa
rs267642	1:57553573	5.28E-06	0.483	0.565	AC	C	1.03 (1.01-1.04)	1.18 (1.00-1.40) [#]	14.55 (1.92-110.03) ^{#*}	F: 2151 M: 2164	215 262	0 15
rs17553946	2:155055407	9.18E-06	0.589	0.967	AG	A	1.59 (1.24-2.03) [#]	1.10 (1.02-1.18) [#]	1.10 (1.05-1.15)	F: 96 M: 157	796 899	1472 1378
rs1568946	7:48338344	7.76E-06	0.526	0.917	AG	G	1.14 (1.08-1.21)	1.17 (1.10-1.26) [#]	1.00 (0.83-1.20) [#]	F: 1258 M: 1135	905 1096	207 214
rs8011192	14:27858369	2.07E-06	0.331	0.656	AC	A	1.38 (1.22-1.56) [#]	1.06 (1.00-1.12)	1.09 (1.01-1.18)	F: 348 M: 495	1175 1149	846 800
rs11159094	14:74165900	1.52E-07	0.096	1.000	CT	T	1.30 (1.18-1.42) [#]	1.10 (1.03-1.16)	1.13 (1.02-1.26)	F: 576 M: 771	1231 1160	555 507
rs7237	15:82987581	5.93E-06	0.483	0.390	AG	A	1.34 (1.09-1.66) [#]	1.13 (1.05-1.22) [#]	1.13 (1.07-1.19)	F: 142 M: 197	846 987	1381 1261
rs11079155	17:50646506	6.50E-06	0.483	0.371	AG	A	1.57 (1.24-1.99) [#]	1.09 (1.01-1.17) [#]	1.11 (1.05-1.16)	F: 106 M: 172	850 953	1407 1312
rs7505612	18:4766396	4.35E-06	0.482	0.853	AC	A	1.57 (1.27-1.94) [#]	1.14 (1.06-1.22)	1.03 (0.98-1.09) [#]	F: 128 M: 207	975 886	1267 1353

Chr, chromosome; HWE, Hardy-Weinberg equilibrium, here P values for HWE were calculated for the combined sexes; PR, proportion ratio, which is the ratio of the proportions of a specified genotype between males and females, with corresponding reciprocal presented if less than one, and with [#] indicating male advantage; CI, confidence interval; Obs, observed.

Positions are in NCBI build-36 coordinates. * Here the number of genotype aa was taken to be 1 in females, where no aa was observed.

The statistical power was calculated according to Purcell et al. [130], with a general 2 d.f. model where the three genotypes at a SNP were assumed to be under different selection pressures between the sexes. For the eight independent SNPs with the strongest signals, all of the values of the proportion ratio were less than 1.6 (except rs267642 where no C/C genotype was observed in females, see Table 3.3), revealing the modest effect of potential SA selection.

As mentioned in Chapter 2.2.3.2, the SNPs failing the WTCCC's HWE filter were excluded to produce high-quality data, although SNPs putatively under natural selection with sex-different effect were likely to violate the HWE. The 3,533 SNPs which failed the HWE filter but passed all other filters were re-examined, and none of them reached $P < 1\text{E-}5$, the threshold used in the genome-wide scan.

In the eight genomic regions of strongest signature, candidate genes, which might be undergoing SA selection, were selected based on three criteria:

- 1) Genes associated with diseases. In theory, sex difference of viability taking place at any stage of lifetime (from the phases of gamete and zygote to adult) prior to sampling can lead to the observed differences in genotype distributions in the study. Hence, genes associated with diseases that exhibit different incidences between the sexes are candidates. However, for many known diseases, there is little information in terms of sex difference. So genes associated with any diseases were presented as candidates in this study. The Online Mendelian Inheritance in Man database (OMIM) [75], which is the most exhaustive public database for

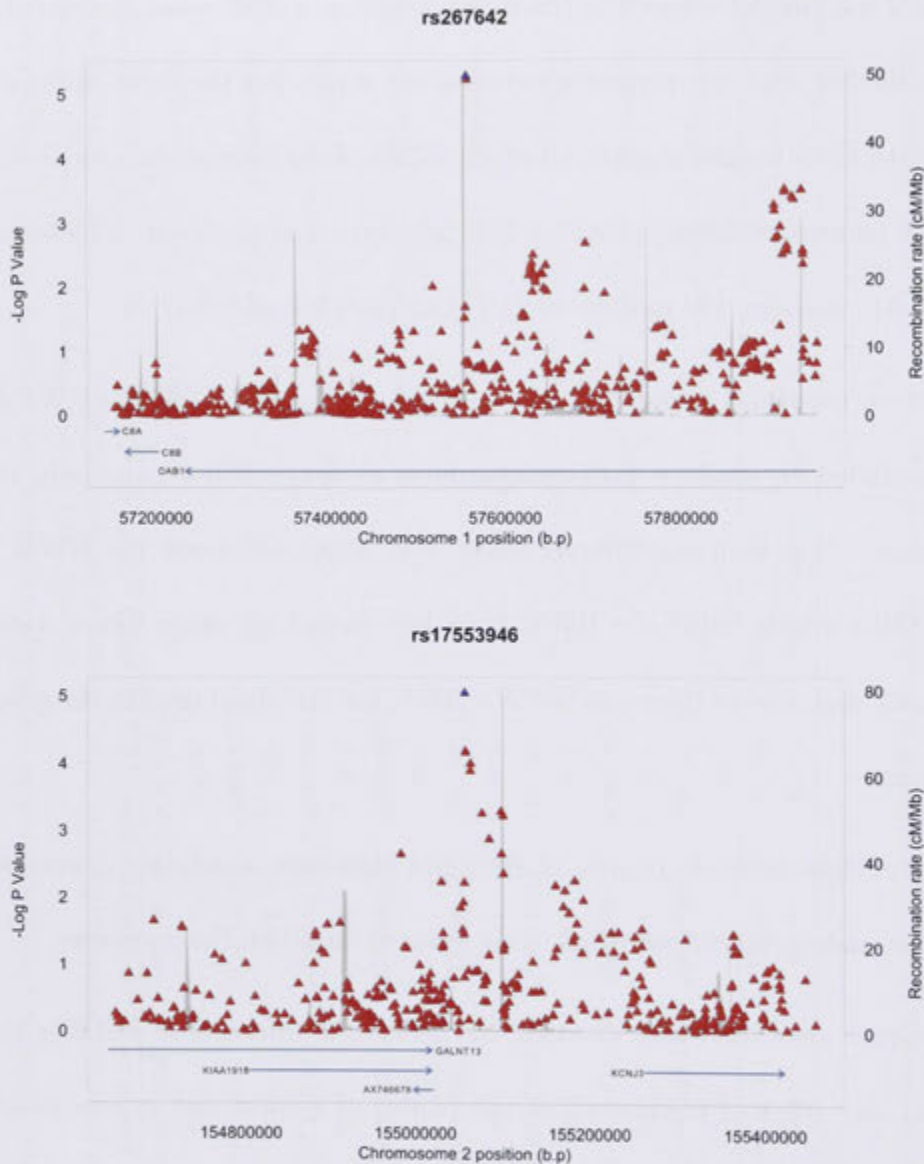


Figure 3.6: The eight genomic regions that reached $P < 1E-5$ in the genome-wide scan. X-axis is genomic location, in NCBI-36 coordinates; y-axis, left, is $-\log(P \text{ values})$, and right, is recombination rate. Triangles represent single SNPs typed in the study, with blue for the peak one in the middle. The grey bars represent recombination rates. In the lower panel, blue arrows are the genes in this region, and gene annotations were taken from the UCSC Genome Browser.

Figure 3.6 continued

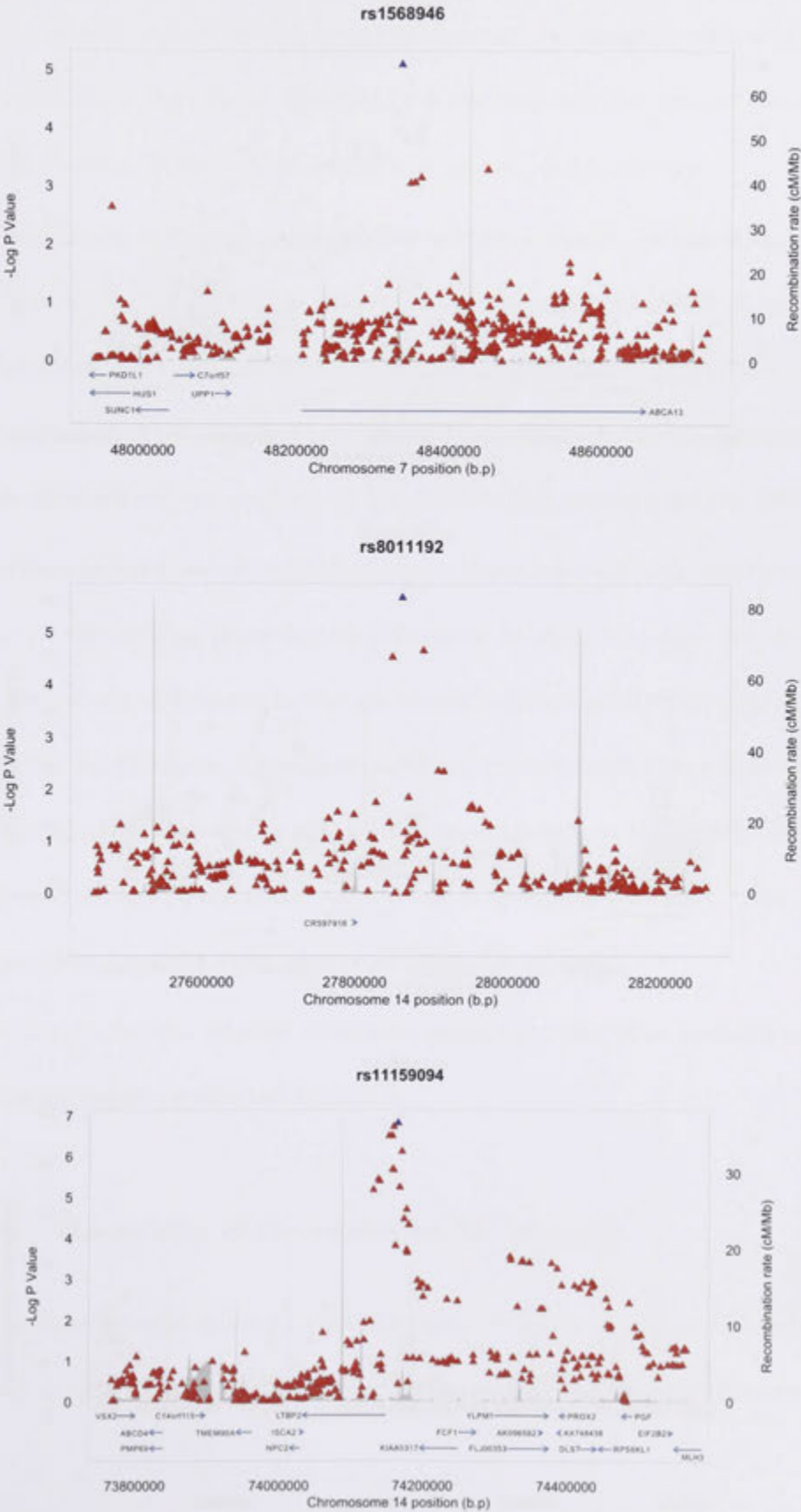
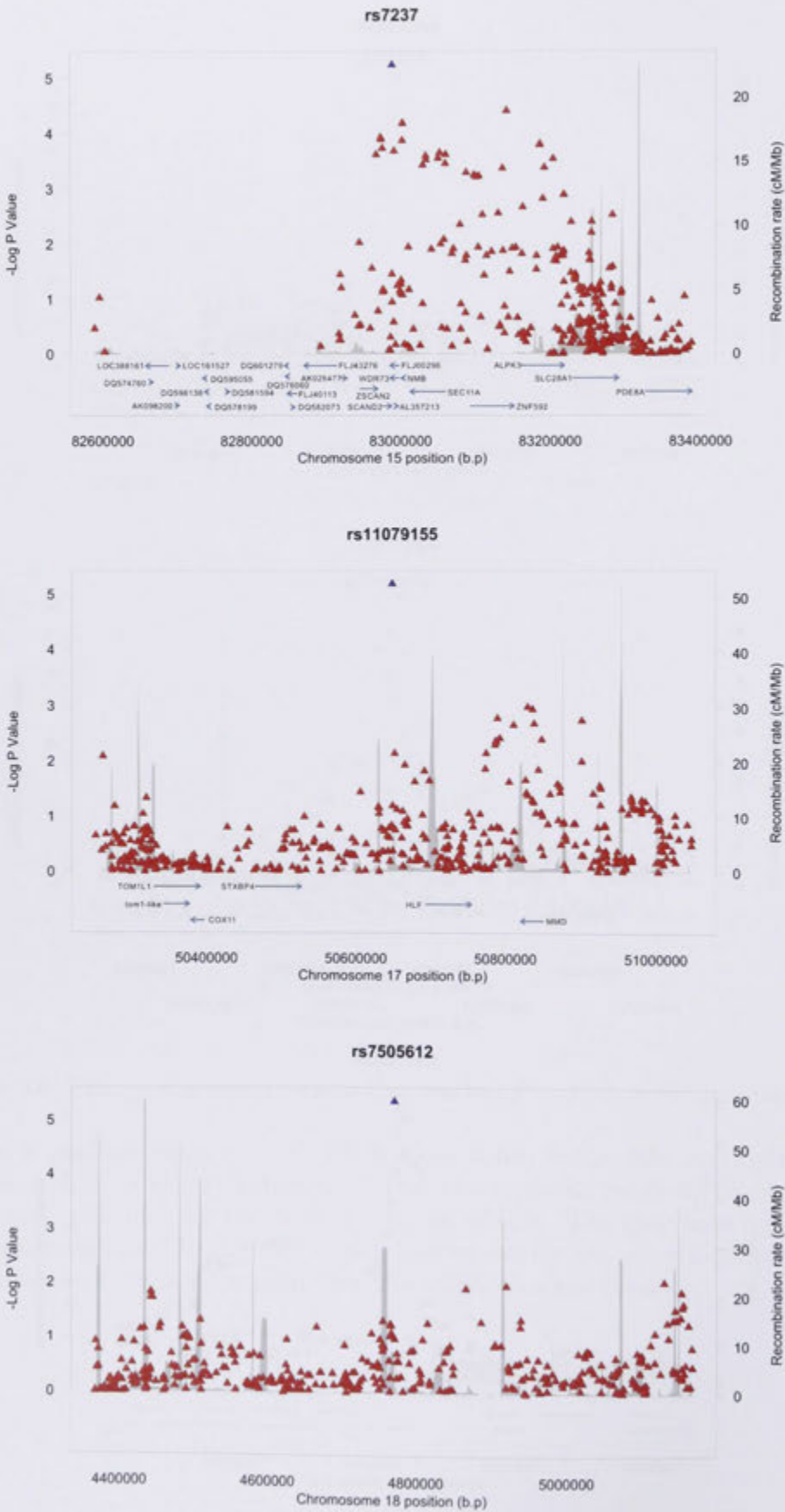


Figure 3.6 continued



the mapping between known Mendelian disorders and genes, was employed. The states for disease onset were also presented once the disease genes were selected as candidates, since analyses in Chapter 4.1.2 revealed that the disease genes with prenatal or infantile onset were enriched by signals of SA selection.

2) Genes with abnormal or lethal effect in knockout mice. Not all human genes are known in respect of disease association, so the reported effects of the genes from knockout mice were considered. Genes were selected as candidates if they led to abnormal or lethal outcomes in knockout-mouse experiments, assuming the relative information was available in the Jackson Laboratory database [85].

3) Genes of functions involving hormones. Hormones, especially sex-hormones, are one of the striking physiological differences between the sexes, which could contribute to sex differences in the prevalence, course and severity of many diseases [173]. Furthermore, the analysis in Chapter 4.2 revealed that genes from the hormone-relative categories by their functions, according to the PANTHER [158] and Gene Ontology [8] databases, were enriched by signals for SA selection. However, no genes from the eight regions satisfied this criterion.

For details for the selected candidate genes from the eight genomic regions with the strongest signals, see Table 3.4.

3.2.2.2 The validity of the signals for SA selection

There was substantial evidence that the eight strongest signals for SA selection revealed in this study were valid. Firstly, the availability of dense genome-wide

Table 3.4: Summary of candidate genes from the eight genomic regions showing $P < 1E-5$ for SA selection

Peak SNP with $P < 1E5$	Candidate genes	Reported OMIM disease	Prenatal or infantile onset	Knockout mouse phenotype
rs267642	DAB1 (disabled, Drosophila, homolog of, 1)	—		Homozygous null mutants show high prenatal and perinatal mortality, with abnormal brain morphology, thymus atrophy and decreased body size.
rs17553946	KCNJ3 (potassium channel, inwardly rectifying, subfamily J, member 3)	—		—
rs1568946	—	—		—
rs8011192	—	—		—
rs11159094	LTBP2 (latent transforming growth factor-beta-binding protein 2)	Glaucoma 3, primary congenital, D	Yes	Homozygous null mutants show embryonic lethality before somite formation.
rs7237	PDE8A (phosphodiesterase 8A)	—		Homozygous null mutants show abnormal hormone level.
	ZNF592 (zinc finger protein 592)	Spinocerebellar ataxia, autosomal recessive 5	Yes	—
rs11079155	HLF (hepatic leukemia factor)	—		Mutants show dramatically shortened life span, with lethal seizures.
rs7505612*	—	—		—

The candidate genes were selected from the regions shown in Figure 3.6, based on three criteria described in text. The information for knockout mouse phenotype was from Jackson Laboratory database [85].

*No genes are included in the regions of this SNP.

Table 3.5: Surrounding SNPs of the peak from the eight genomic regions showing $P < 1E-5$ for SA selection

Peak SNPs ^a	No. of selected surrounding SNPs with $P < 0.05$ / No. of all selected surrounding SNPs ^b
rs267642	2 / 3
rs17553946	9 / 12
rs1568946	3 / 10
rs8011192	14 / 29
rs11159094	40 / 49
rs7237	14 / 35
rs11079155	2 / 2
rs7505612	0 / 2 ^c

^aFigure 3.6 shows signals from the SNPs around these eight peak SNPs.

^bOnly surrounding SNPs in linkage disequilibrium (LD) with the peak SNP were selected. A SNP will be considered as in LD with the peak one if the corresponding LD statistic $r^2 > 0.2$. The LD data were from the HapMap project.

^crs7505612 had two linked SNPs, rs12454604 ($r^2 = 0.458$) and rs8089250 ($r^2 = 0.414$), which exhibited boundary P values for SA selection, 0.055 and 0.059 respectively.

Table 3.6: P values at the eight SNPs with the strongest signals for the two control cohorts

SNP	P of combined controls	P of 58C	P of UKBS
rs267642	5.28E-06	0.033	5.13E-05
rs17553946	9.18E-06	0.006	0.001
rs1568946	7.76E-06	2.17E-04	0.020
rs8011192	2.07E-06	9.25E-05	0.011
rs11159094	1.52E-07	3.22E-04	4.88E-04
rs7237	5.93E-06	0.008	5.01E-04
rs11079155	6.50E-06	0.002	0.002
rs7505612	4.35E-06	0.001	0.003

58C, the 1958 British Birth Cohort; UKBS, UK Blood Services.

SNP data allows me to be confident that these results are not due to genotyping errors. For the eight SNPs, the surrounding SNPs also showed significant signals, which was consistent with the linkage relationship between these SNPs (see Table3.5 and Figure3.6). Since the SNPs genotyped on two different platforms were combined, the signals from the two platforms were checked respectively, and their consistency is shown in Figure 3.7, confirming no platform bias in the data used in this study. For the eight SNPs, significant signals were showed in each of the two controls, with stronger signal in combined controls, which exhibited the concordance of genotyping outcome between the two control cohorts (see Table 3.6).

Secondly, there was no evidence that the strongest signals were caused by the confounding genotyping of genomic duplication between the autosomal and sexual chromosomes. The quality control against duplicate SNPs was done before conducting the genome-wide scan (see Chapter 2.2.3.3). The deviations from HWE are good indicators for duplicating SNPs, and produce a power of 80% for

sample sizes as small as 50 [92]. However, all P values for HWE test at the eight SNPs were greater than 0.37, confirming these results were not due to duplication (Table 3.3).

Thirdly, population stratification is unlikely to be responsible for the results in this study. Population subgroups are defined by ethnicity or geographic origin and consequently differ in allele frequency [76]. If population subgroups did not equally affect males and females, artificial positives would appear. The common statistic used in genome-wide studies to assess the effect of population stratification is genomic control inflation factor, which is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median [43]. The closer values are to 1 the weaker the effect. The genomic control inflation factor in this study was 1.011, indicative of an extremely weak effect. Moreover, the 13 ancestry informative markers described by the WTCCC, which vary substantially in allele frequency across the UK [170], did not show significant signals during the genome-wide scan (all P values > 0.096).

Lastly, sex differences in mutation rate may lead to positives in the study, but the mutation rate is extremely low in humans. With an estimated de novo germline mutation rate of $1\text{E-}8$ per base pair per generation [1], the expected number of mutation was less than one in the 1.2 million SNPs of this study.

3.3 Discussion

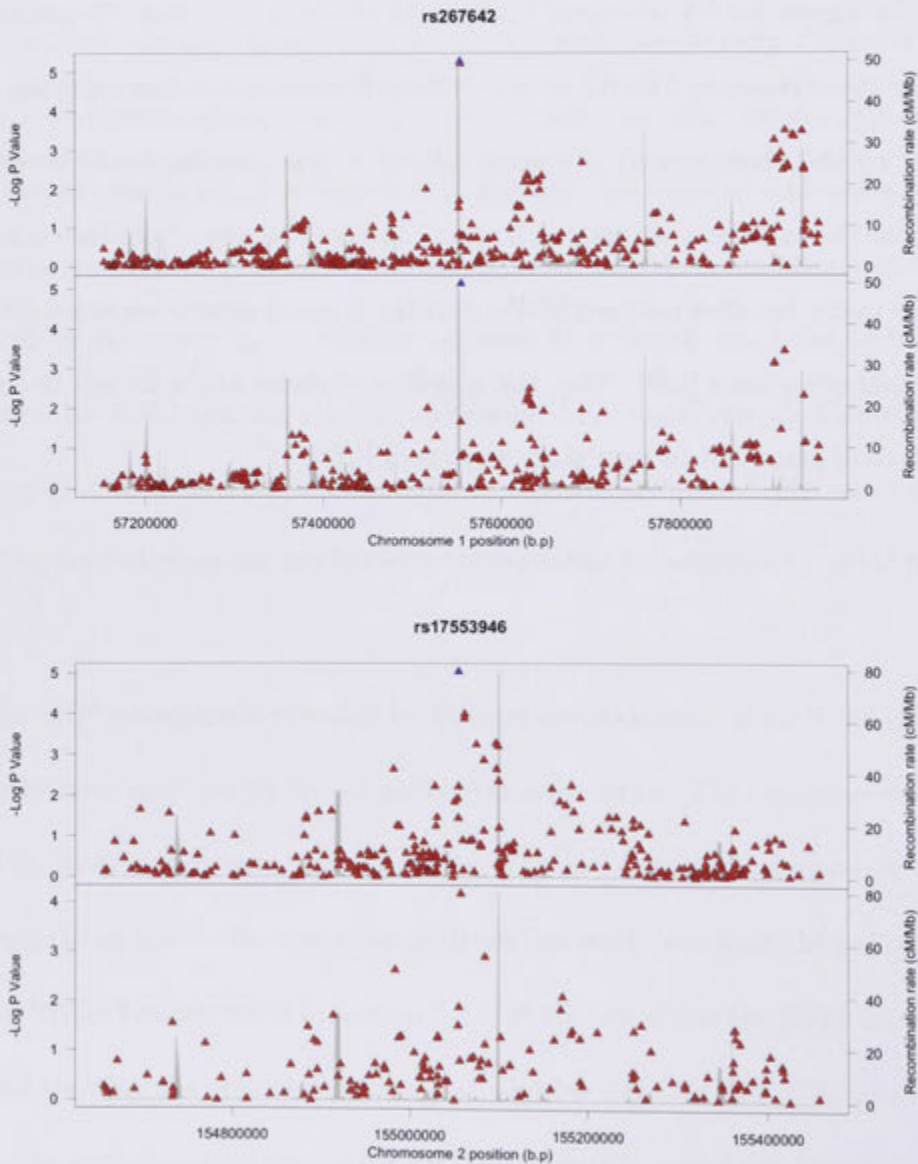
From an evolutionary perspective, in general, the viability selection acting on modern human population has been greatly weakened through contemporary medical technology and human modification of the environment [148]. This may lead to the low power to detect SA selection in humans. In this chapter, the observed test statistics from the genome-wide scan significantly departed from the expected values (shown in the Q-Q plot Figure 3.4). This result supports the existence of genetic sexual antagonism in humans. Substantial evidence was also presented to support the validity of the signals. However, the strongest signal revealed by the study ($q = 0.096$ or adjusted $P = 0.080$) was close to but did not reach the 0.05 level, the traditional threshold for ‘significance’. This suggests that replication in independent samples is needed, and that bigger sample size might be needed given the modest effect of potential SA selection revealed by the study.

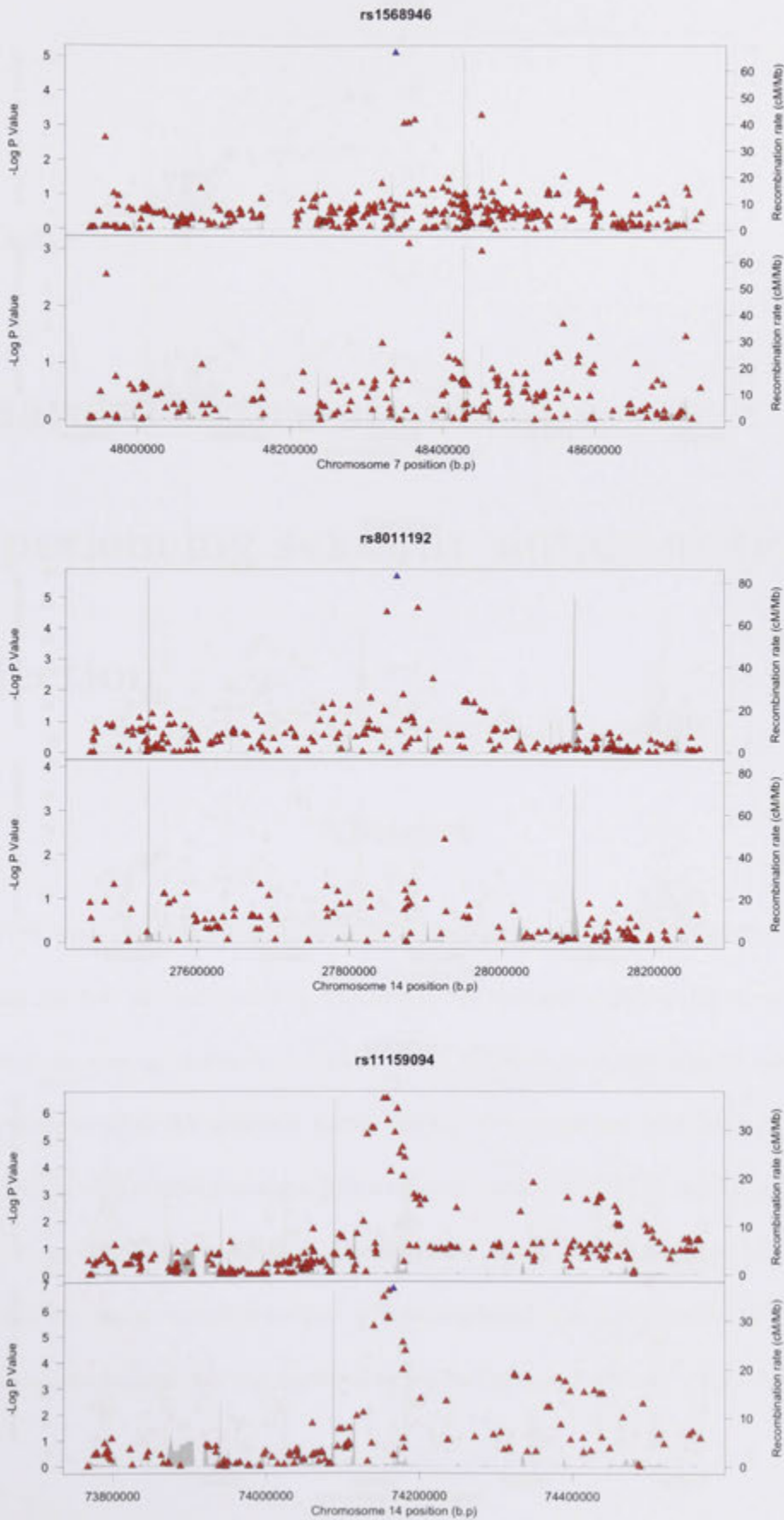
The SNP associations revealed by the genome-wide scan, even if the associations are true, may not be the causative variants per se. The causative variants could be those which were not determined but in LD to the surrounding SNPs interrogated in the study. One example from this study was from the genomic region of BRCA2 as discussed in section 3.2.1. Although it was the SNPs located in BRCA2 showing the observed peak signal, another gene, LOC646799, zygote arrest 1-like gene (Ensembl gene ID: ENSG00000189167), could not be ruled out as the causative gene. LOC646799 has the peak signal located at its upstream with

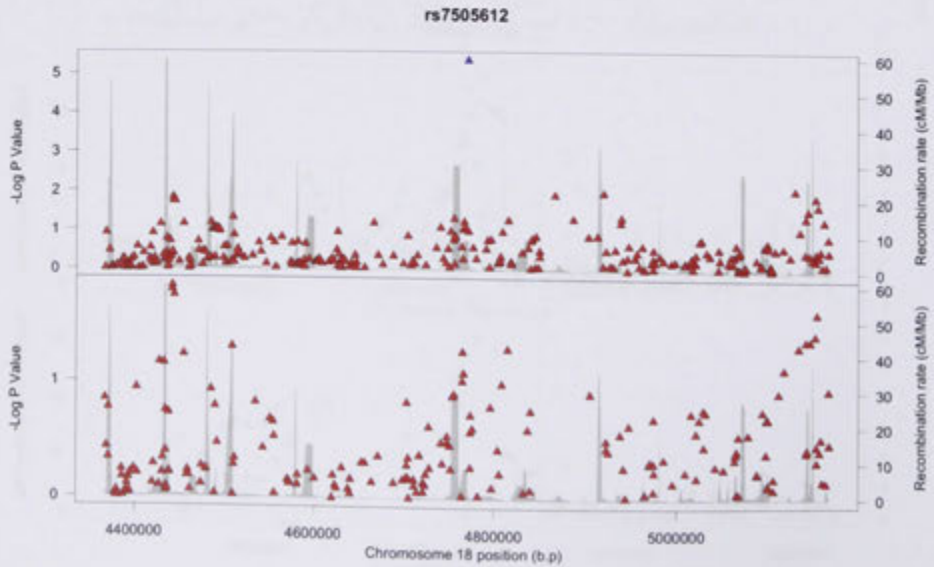
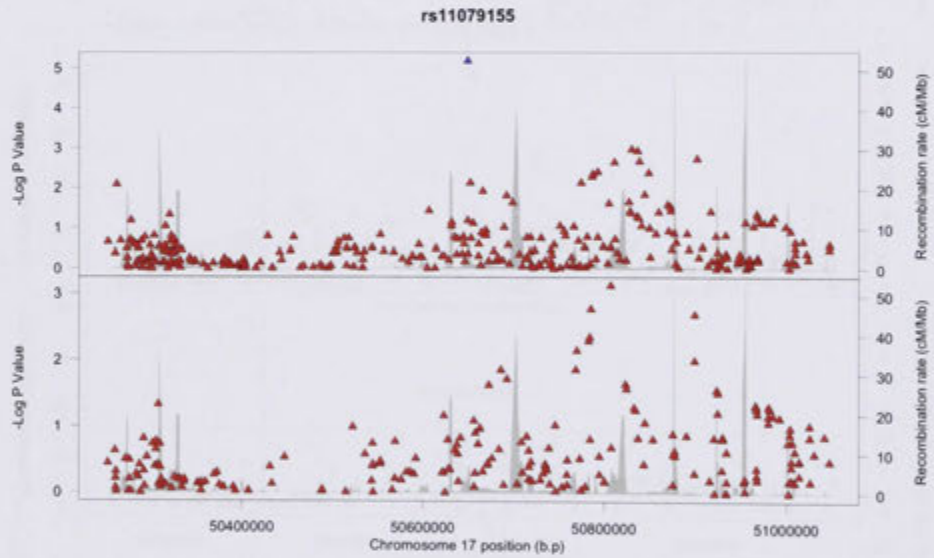
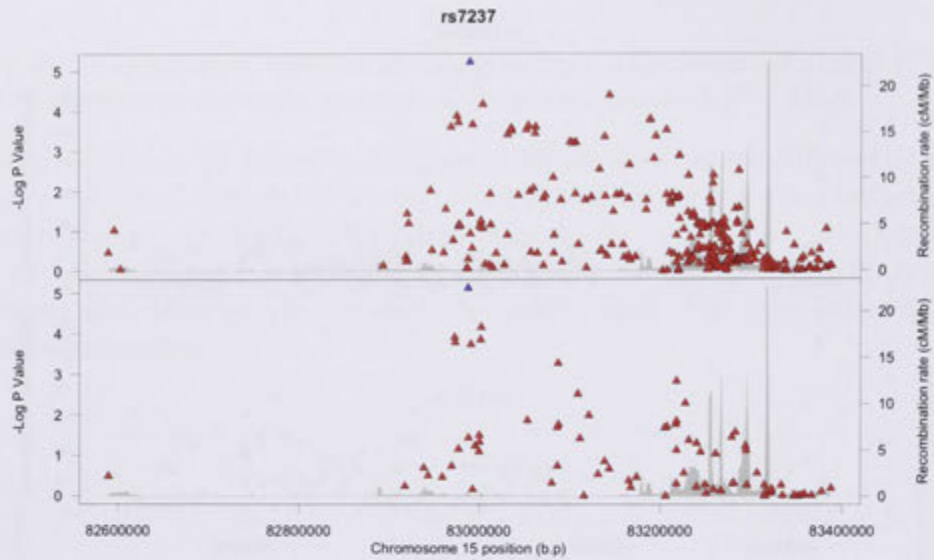
25-kb distance (Figure 3.2). Moreover, LOC646799 is predominantly expressed in oocytes and early embryos, and may function as an RNA regulator, according to the description from the NCBI RefSeq project (RefSeq ID NM_001136571, at http://www.ncbi.nlm.nih.gov/nuccore/nm_001136571). Given the discovery in Chapter 4.1.2 that the disease genes with prenatal or infantile onset were enriched by signals for SA selection, LOC646799 deserves attention. To determine whether the variants in BRCA2 or in LOC646799 is causative, fine-mapping (e.g. sequencing the whole region), is needed. Likewise, fine-mapping is also needed to determine the causative variants in all other candidate regions. Note the causative variants might be other untyped SNPs, or other types of genetic variation such as copy number variants [100]. Also, the causative variants might be not in coding sequences of genes but in regulatory sequences [122].

Figure 3.7: Comparison between the signals from Affymetrix V6.0 and Illumina 1.2M platforms, at the eight genomic regions that reached $P < 1E-5$.

The upper panel is for Illumina platform, and the lower is for Affymetrix. Note that only $\sim 18\%$ of the total SNPs were genotyped on both of the platforms. X-axis is genomic location, in NCBI-36 coordinates; y-axis, left, is $-\log(P \text{ values})$, and right, is recombination rate. Triangles represent single SNPs typed in the study, with blue for the one with the peak signal. The grey bars represent recombination rates.







Chapter 4

Characteristics of variants experiencing sexually antagonistic selection

Abstract

SNPs with very high density were employed in this study. Between the SNPs passing the QC and analyzed in the study, the median physical distance is 1.2 Kb, and the average distance is 2.2 Kb. With SNPs of such high density, the signal distribution for SA selection across the whole genome has been estimated by tests for sex-differential genotype frequencies at each of the SNPs, in the previous chapter. In this chapter, whether specific genes or genomic regions are relatively enriched for signals of SA selection will be explored and determined. With the available annotations for the enriched genes and genomic regions, characteristics

of the variants experiencing SA selection can be revealed, and several theories and hypotheses regarding SA selection can be tested. Firstly, the hypothesis, that it is the sex difference in viability or mortality leading to the observation of differences in genotype distributions between the sexes, was tested by the comparison between controls and cases at the genomic regions with strongest signals for SA selection, and by the examination of whether there was signal enrichment in the group of genes reportedly associated with diseases. All the results support this hypothesis. Secondly, signals from the genes associated with sex hormones, one of the major physiological differences between the sexes, were examined. The results suggest these genes are hotspot of ongoing SA selection. Thirdly, the evolutionary hypotheses, that the evolutions of sex-specific cis-regulatory elements, imprinted genes and positively selected genomic regions could be driven by SA selection, were tested, but no positive results were obtained. Lastly, evidence was presented that, due to SA selection, the differences, not only in genotype frequencies, but also in allelic frequencies, can be observed between the sexes in human random samples.

4.1 Human diseases underlying SA selection

In the previous Chapter 3, evidence for SA selection in humans had been presented in the form of observed differences in autosomal genotype frequencies between the sexes. In theory, sex differences in viability or mortality could lead to such differences, because males and females share the same probability of inheriting the

same autosomal genotypes from their parents. Hence, the genes associated with sex-specific diseases were assumed to be enriched by signals of SA selection. In this section, the relationship between human diseases and SA selection is explored.

4.1.1 Sex-effects of diseases underlying SA selection

In the genome-wide scan described in Chapter 3, eight genomic regions with strongest signals for SA selection were presented (see Table 3.3). Besides the data for control individuals used in the genome-wide scan, SNP data for seven disease cohorts were also employed in this study, as described in Chapter 2.1.1. At the eight significant genomic regions, the comparison between the controls and cases in each sex could reveal whether there were sex-effects from the diseases, which would help to explain the observation of sex differences in genotype distributions among the control samples.

However, the case individuals were genotyped on different platform by the WTCCC, with only $\sim 30\%$ of the SNPs genotyped in controls also available for the cases. Only two of the eight peak SNPs (rs267642 and rs7237), exhibiting the strongest signal in the eight presented genomic regions, were also genotyped in the case cohorts. For the other six peak SNPs, surrogates were sought for case-control comparison. Surrogate SNP was defined as a SNP, which was genotyped in both cases and controls, and located closely to and in the strongest linkage disequilibrium (LD) with the peak SNP. The extent of LD between the peak and the surrounding SNPs was measured by r^2 sourced from the HapMap project [76].

The surrogate SNP was selected only when the corresponding $r^2 > 0.2$, since surrounding SNPs with weak LD relationship may not be representative of SNPs with the strongest observed signal. For this reason, three peak SNPs did not have surrogates available. For the other three peak SNPs, surrogate SNPs were determined, and they were rs1150137 ($r^2 = 0.29$) for rs17553946, rs8008977 ($r^2 = 0.359$) for rs8011192, and rs1558933 ($r^2=1.0$) for rs11159094. The r^2 of 1.0 for rs1558933 means complete LD between the peak SNP and its surrogate. These three surrogate SNPs, plus the two peak SNPs originally genotyped both in cases and controls, were selected for case-control comparison.

The comparison of genotype distribution between cases and controls was conducted in each sex. There were a total of 70 tests for case-control comparison at the five selected SNPs, among which seventeen showed positive results, with $P < 0.05$ as the cut-off threshold (see Figure 4.1). The proportion of positives (17/70) was significantly high ($P = 0.006$, one-sided Fisher's exact test), compared to ~ 4 among the 70 tests under the null model of no sex-differential effects between cases and controls. For any disease, all the five SNPs had effects only in one sex, except rs1150137 on hypertension. At the five SNPs, genotype distributions were also compared between the sexes in the disease cohorts, with four signals with $P < 0.05$ revealed. These four signals were rs1150137 for hypertension ($P = 0.033$), rs7802110 for coronary artery disease ($P = 0.0006$), rs1558933 for Crohn's disease ($P = 0.003$), and rs12450116 for rheumatoid arthritis ($P = 0.002$).

Sex-specific effects have been reported for some of the seven diseases repre-

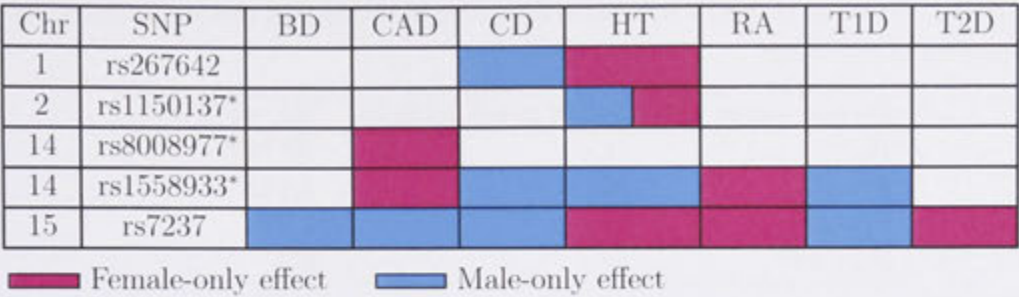


Figure 4.1: Comparisons of genotype distribution between cases and controls in each sex, at the genomic regions that exhibited strongest signals for SA selection during genome-wide scan. Colored cells represent the comparison results with $P < 0.05$, with pink for females and blue for males. SNPs marked with * are surrogates for the peak SNP of the strongest signal within the region. The selection criteria for surrogate SNPs are described in the text. BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn’s disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes.

sented by the case cohorts. For example, there is a striking imbalance between the sexes in the occurrence of rheumatoid arthritis, with females representing the majority of the cases [3,113,114,174]. For cases of type 1 diabetes, male excess is a consistent finding in populations of European origin aged 15-40 years [62]. Large studies of Crohn’s disease demonstrated greater female prevalence than males, in North America, Scandinavia and Northern France [21]. As for coronary artery disease, women experience higher morbidity and mortality [47]. Hence, the eight genomic regions with strongest signals for SA selection discovered by this study were good candidates for studying the sex-effects of these diseases. In Figure 4.1, three of the five selected SNPs were sexually associated with more than one diseases, which might suggest that multiple diseases with sex-differential effect, probably including more other diseases unknown or not included in this study, caused the observation of sex differences in genotype distributions among

Table 4.1: Number of SNPs located within the hOMIM [16] disease genes of four categories of onset age

Disease-onset age	No. of genes included	No. of SNPs
Birth	585	20,789
Pre-15	262	8,182
Pre-40	59	1,978
Post-40	14	471

the control samples.

4.1.2 Signal concentration of diseases genes

If sex differences in viability or mortality led to the differences in genotype distributions observed in this study, the signals for SA selection are likely to be concentrated in genes associated with diseases with sex-specific effect. For most diseases, there is little information about sex effects, but nearly all human diseases are thought to have more or less sex-specific differences in prevalence, age of onset and/or severity [112]. So regardless of the awareness of their sex effects, whether the reported disease genes were enriched by signals for SA selection is the main question to be answered below.

The Online Mendelian Inheritance in Man database (OMIM) [75] and a hand-curated version of OMIM [16] (hereafter “hOMIM”) were employed to define the disease genes. There are ~2,500 autosomal genes mapped to disorders in OMIM, up to March 2011. However, the phenotypic information from OMIM is sometimes outdated and is manually edited without a standard format [16]. So hOMIM, a hand-curated version of OMIM by Blekhman et al. [16] to include only highly penetrant diseases caused by point mutations, was also employed in this study.

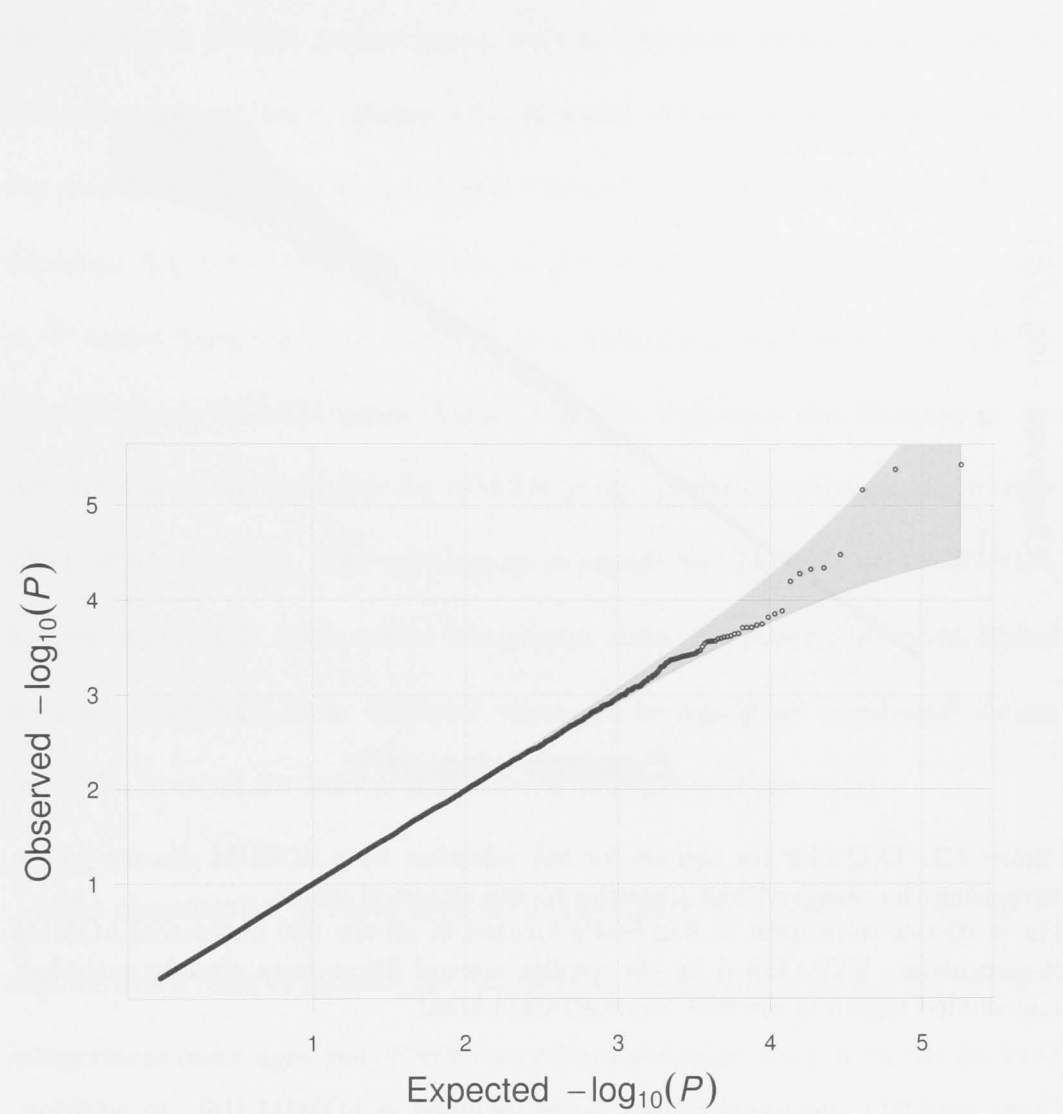


Figure 4.2: Q-Q plot for signals for SA selection from OMIM disease genes, comparing the observed test statistics to the expected values. The statistics were from 102,246 SNPs located in all the 2,496 autosomal OMIM disease genes. SNPs failed by the quality control filters were already excluded. The shaded region is the 95% concentration band.

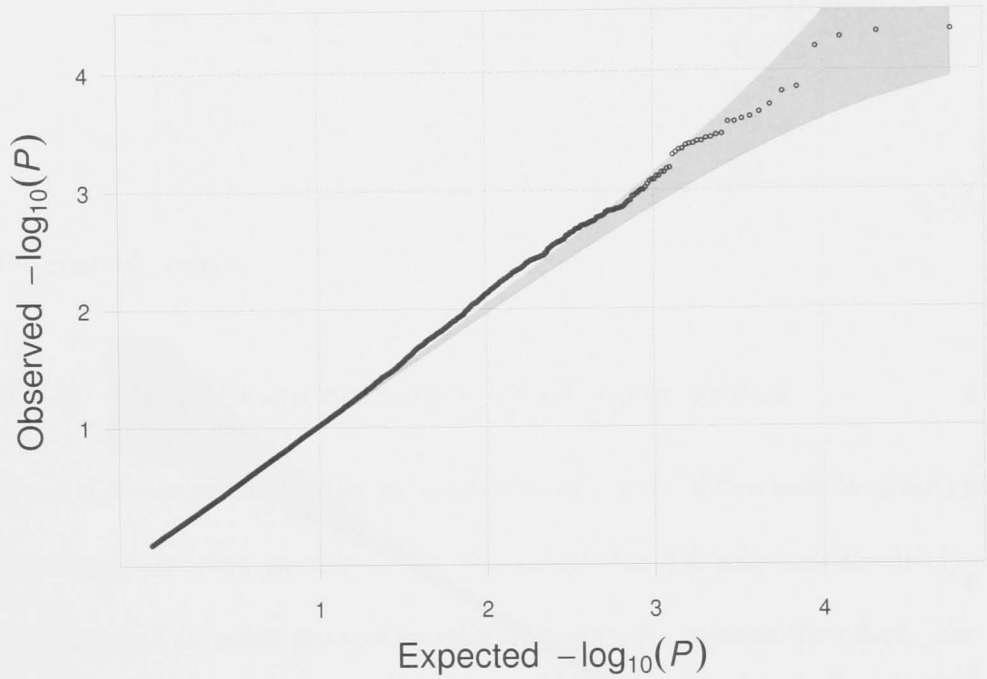


Figure 4.3: Q-Q plot for signals for SA selection from hOMIM disease genes, comparing the observed test statistics to the expected values. The statistics were from 31,942 SNPs located in all the 939 autosomal hOMIM disease genes. SNPs failed by the quality control filters were already excluded. The shaded region is the 95% concentration band.

There were 939 autosomal disease genes included in hOMIM [16]. In addition, hOMIM classified the ages of disease onset into four categories, “in utero or shortly after birth”, “before 15 years of age”, “between 15 and 40 years of age”, and “post-40”, which would be also utilized in this study (See Table 4.1 for the numbers of SNPs within genes of the four categories).

The signals for SA selection from the SNPs located in disease genes were compared to their expected distribution under null condition, via Q-Q plots. The comparison showed no significant enrichment for the 102,246 SNPs located in all

the autosomal OMIM disease genes, with no observed outliers apart from the 95% concentration band (Figure 4.2). However, for the 31,942 SNPs located in the disease genes included in hOMIM dataset, the comparison showed obvious departure for the observed statistics in the range of ~ 7 to 12 (corresponding to P values from ~ 0.03 to 0.002, 2 d.f.), indicating enrichment of signals for SA selection in hOMIM genes (Figure 4.3). No departure was observed at the extreme tail in the Q-Q plot for hOMIM genes, probably reflecting the modest effect of SA selection. The enrichment of signals for SA selection on hOMIM, but not on OMIM, might reflect the greater statistical power due to the higher accuracy of hOMIM data. Similarly, there was no significant enrichment among the SNPs from all the known human autosomal genes (Figure 4.4).

The classification of onset ages for disease genes by hOMIM can be used to explore which class of diseases the enrichment of signals were from. For the four categories of onset ages, only SNPs located in the disease genes from the group of ‘in utero or shortly after birth’ showed significant enrichment via Q-Q plots (see Figure 4.5). The trend that signals for SA selection are weaker with the increase of disease onset age was shown, when signals from SNPs of the four categories of disease-onset age were plotted on one figure (see Figure 4.6). Such trend is also statistically supported by the comparisons of the distribution of these signals with one-sided Kolmogorov–Smirnov test between the Birth and Pre-15, between the Pre-15 and Pre-40, and between the Pre-40 and Post-40 groups, which generated P values of 0.021, $3.5\text{E-}10$ and $8.1\text{E-}4$, respectively. The significant enrichment of

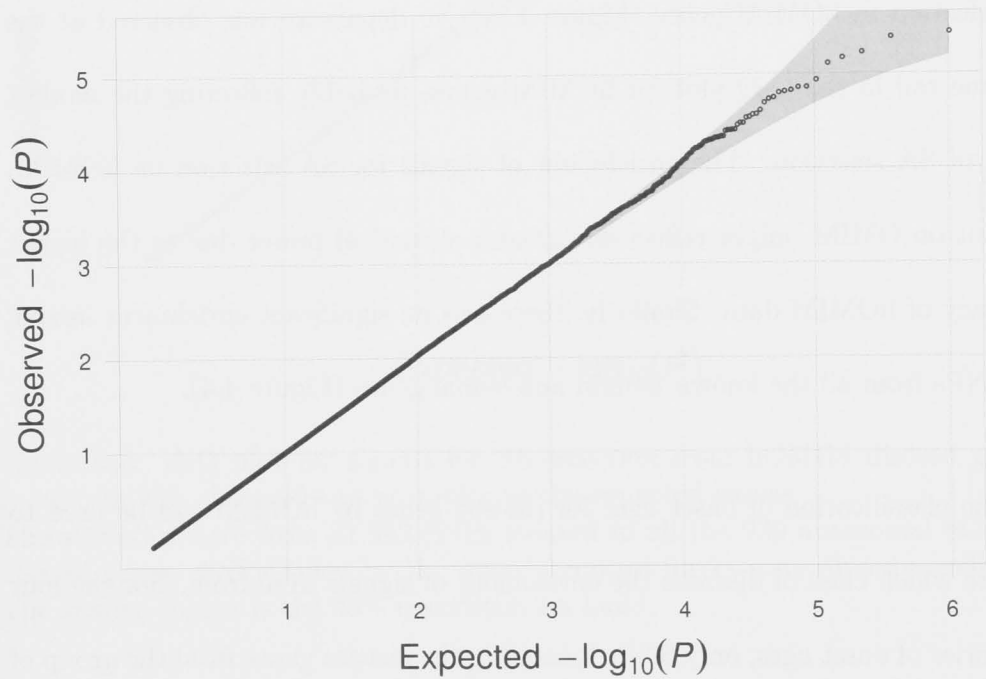


Figure 4.4: Q-Q plot for signals for SA selection from all known autosomal genes, comparing the observed test statistics to the expected values. The statistics were from 571,791 SNPs, located in all the 24,252 known autosomal genes according to Ensembl database release 59 [129]. SNPs failed by the quality control filters were already excluded. The shaded region is the 95% concentration band.

signals from the disease genes with prenatal or infantile onset is consistent with the different infant mortalities between the sexes in humans, with the ratio of males to females ranging from ~ 1.25 to 1.30 in the UK during 1950-2000 [46]. The trend that signals for SA selection are weaker with the increase of disease onset age reflects the changes of statistical power, because the earlier the disease onset is, the greater proportion of individuals are affected by the disease.

4.2 Functional category enrichment

This section explored what function categories of genes are enriched by signals for SA selection. The gene classification by function was from PANTHER [158]. PANTHER (version 6.1) classified human genes into 242 terms, in terms of biological process (such as pathways, cellular roles or physiological functions). To measure the degree of significant signal enrichment for a specified gene category, SNPs tested in the genome-wide scan and located in the genes from each gene category were grouped. For the SNPs in each gene category, the proportion of SNPs with $P < 0.01$ for SA selection was calculated and compared to that proportion in all the ~ 1.2 million SNPs used in the genome-wide scan (hereafter the ‘background proportion’), to generate a P value measuring the significant signal enrichment. The P value was calculated by one sided Fisher’s exact test, with that the smaller the P value, the higher the degree of signal enrichment for that gene category. The usage of $P < 0.01$ for SA selection for individual SNPs was to detect the modest effect of potential SA selection, although this signature

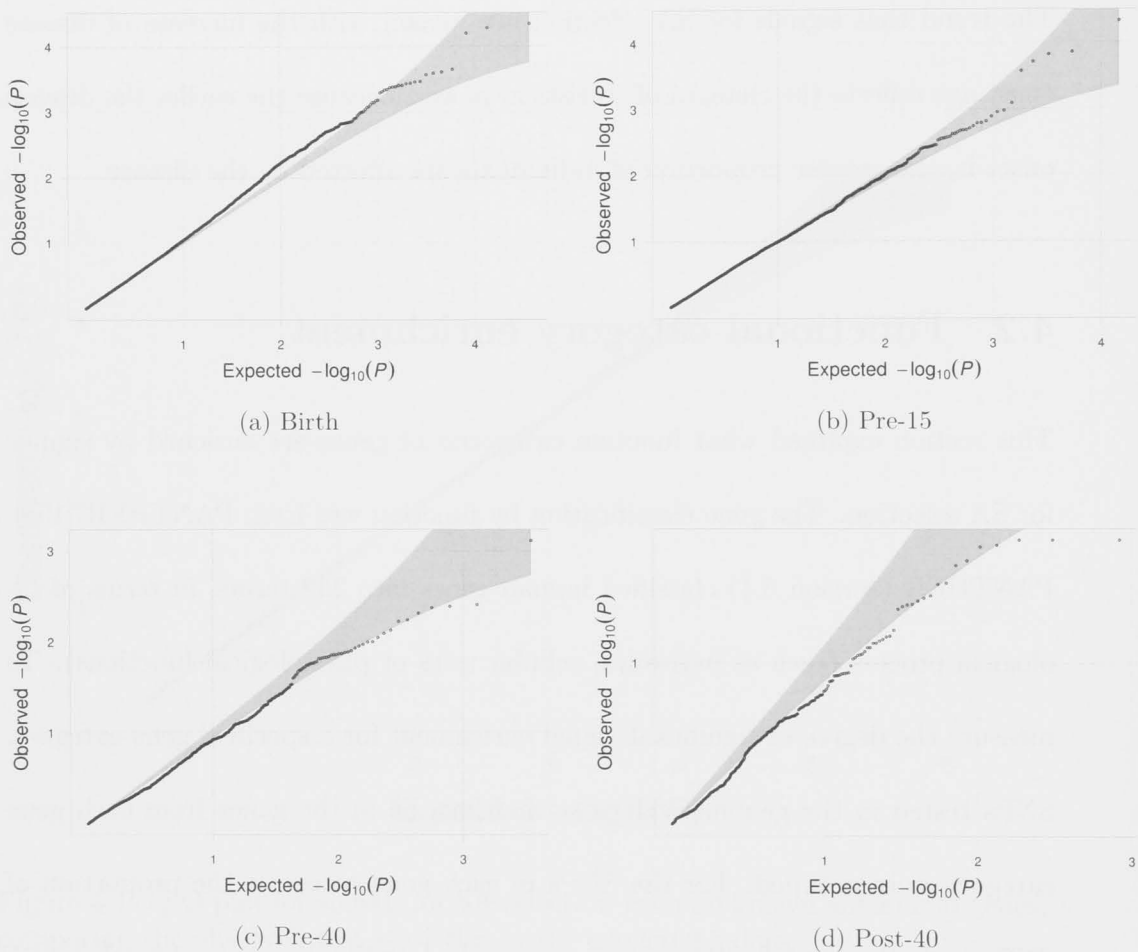


Figure 4.5: Q-Q plots for the signals for SA selection from SNPs located in the disease genes from each of the four categories of disease-onset age. The four categories were “in utero or shortly after birth”, “before 15 years of age”, “between 15 and 40 years of age”, and “post-40”, according to hOMIM [16]. Note in the hOMIM data, a gene may be associated with multiple Mendelian diseases, as well as different onset ages. In that case, the gene was only assigned to the category of the earliest onset age in the study.

Figure 4.6: Comparison of the signals for SA selection from SNPs located in the hOMIM disease genes from four categories of disease-onset age. The diagonal line is $y = x$.

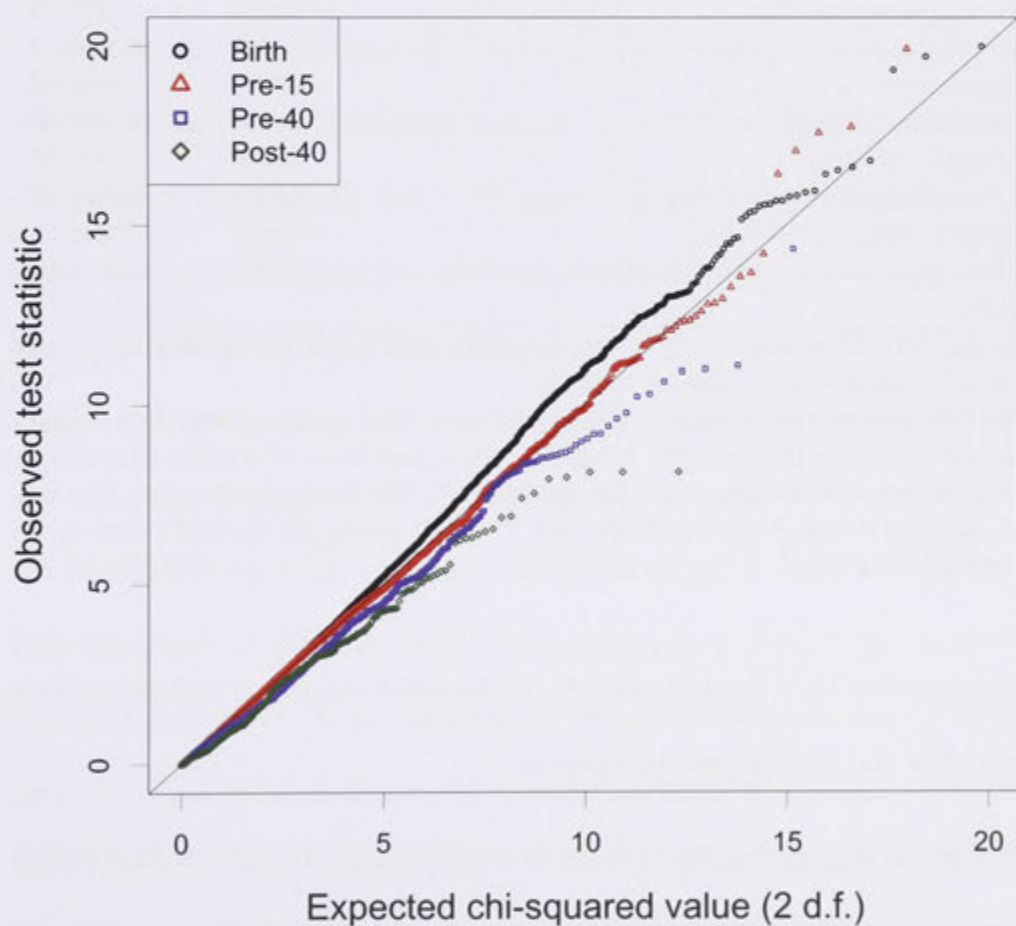


Table 4.2: Enriched PANTHER biological process categories

Biological process	Proportion of positive SNPs [#]	Q value [*]
Steroid hormone metabolism	6.92%	3.68E-18
Electron transport	2.74%	2.58E-13
mRNA polyadenylation	6.50%	2.12E-11
Other lipid, fatty acid and steroid metabolism	4.53%	5.53E-09
Mitosis	2.29%	2.10E-08
MAPKKK cascade	1.92%	2.62E-07
Anion transport	2.20%	1.21E-05
Sulfur metabolism	2.21%	9.41E-05
Other signal transduction	2.48%	2.48E-04
Fatty acid beta-oxidation	2.52%	7.43E-03
Heart development	1.99%	8.09E-03
Gametogenesis	2.82%	8.09E-03
Carbohydrate transport	2.06%	8.09E-03
Phosphate transport	3.28%	2.26E-02
Gut mesoderm development	2.84%	2.82E-02
Anterior/posterior patterning	1.80%	3.23E-02
Receptor protein serine/threonine kinase signaling pathway	1.74%	4.62E-02

[#]Here the SNPs showing $P < 0.01$ for sex-differential genotype distribution in the control individuals were defined as ‘positive’.

^{*}Only 17 categories with q value < 0.05 were listed. The q values were the multiple-testing correction for the total 242 categories, for comparisons of positive proportions between the specified category and the background. The background proportion was 1.06%, and was the proportion of SNPs with $P < 0.01$ among all the SNPs used in the genome-wide scan.

might only reflect the LD relationship to the causative variant, or be false positive generated by the multiple testing problem.

Testing the degree of signal enrichment simultaneously for the 242 PANTHER gene categories raised the multiple testing issue. To solve such issue, q values for false discovery rate were calculated. The 17 categories with q value < 0.05 are listed in Table 4.2. Three of the 17 gene categories, ‘steroid hormone metabolism’, ‘other lipid, fatty acid and steroid metabolism’ and ‘fatty acid beta-oxidation’,

are thought to be among the major physiological differences between males and females [5, 15, 81, 96, 126]. Another five categories, ‘mitosis’, ‘heart development’, ‘gametogenesis’, ‘gut mesoderm development’ and ‘anterior/posterior patterning’, are assumedly related to the diseases with extremely early onset, and corresponded to what was revealed with the hOMIM data above. The degree of signal enrichment for all the 242 PANTHER terms is detailed in Appendices TableA1.

Another database for gene classification, Gene Ontology [8], was employed to confirm the significant enrichment from the genes associated with sex hormones. Compared to PANTHER, Gene Ontology used much more detailed ($\sim 5,000$) terms to describe biological process for gene classification [157], which led to very few or even no SNPs located in the genes from many terms, and consequently was not appropriate for genome-wide analysis. However, Gene Ontology classified genes in different way, and included part of the genes that were not embodied by the PANTHER database. Searching by “sex hormone” in Gene Ontology online (4th September 2010) returned 14 terms of biological process (see Appendices Table A2 for details). From the 14 terms, there were 96 genes, among which 77 genes were not included within the 17 enriched PANTHER terms. 1,961 SNPs were located in these 77 genes, and 1.74% of these SNPs were with $P < 0.01$, indicating significantly high concentration compared to the 1.06% of background proportion ($P=0.005$, one-sided Fisher’s exact test). This was also consistent with the departure from the 95% concentration band showed in the corresponding Q-Q plot (Figure 4.7).

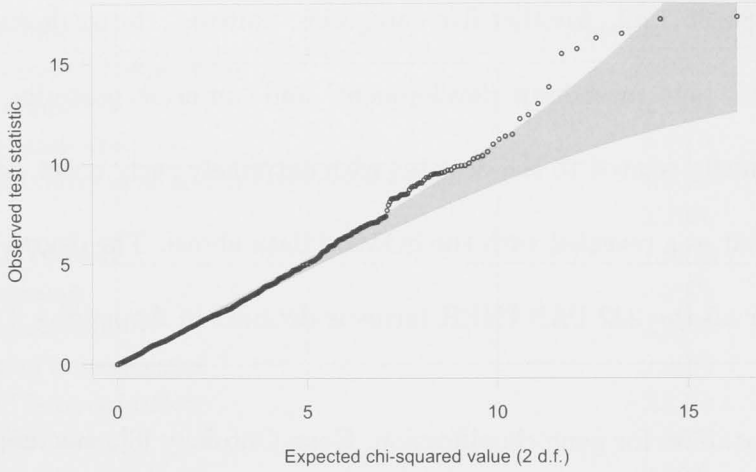


Figure 4.7: Q-Q plot for the 1,961 SNPs located in the 77 genes, which are associated with regarding sex hormones, and are embodied by the Gene Ontology but not by the enriched PANTHER terms.

4.3 No enrichment in sex-hormone receptor binding loci

Recent studies quantifying mRNA abundances using microarrays have suggested that sex-biased gene expression is widespread across organisms and genomes [49]. During sex-biased gene expression, genes are expressed exclusively in one sex, or in both sexes but at a higher level in one sex. Theories suggested that intralocus sexual conflict plays a key role in driving the evolution of sex-biased gene expression [49, 163]. This happens because sex-biased gene expression increases the fitness of the combined sexes in the case of sexual antagonism, by the shared allele expressed solely or more in the sex that it benefits, rather than in the sex that it harms.

4.3. NO ENRICHMENT IN SEX-HORMONE RECEPTOR BINDING LOCI 77

One of the hypothetical mechanisms underlying sex-biased expression is the constant regulatory differences between males and females [112]. The binding between cis-regulatory elements and trans-regulatory elements is a very common way to regulate gene expression [86]. A cis-regulatory element is a region of DNA usually located on the same chromosome with the gene to be regulated, while trans-regulatory elements are usually proteins. A study in mice showed that different cis-regulatory elements might be used between the sexes [13], maintaining the sex-biased expression. The evolution of sex-specific cis-regulatory elements might be driven by SA selection. Furthermore, if sex-specific cis-regulatory elements exist widely across the human genome, then mutations in such regions can impact only one sex, which could result in sex-differential viabilities or mortalities, and consequently the observation of sex-differential genotype distributions from this study. Whether the WTCCC SNPs in cis-regulatory elements are enriched of signals for SA selection will be examined in this section.

Sex-hormone receptor binding loci (SHRBL) are common cis-regulatory elements in humans. SHRBL are the DNA regions which sex-hormone receptors bind to, causing the consequential regulation of gene expression. The SHRBL for two receptors for two major sex-hormones, estrogen and androgen, were considered in the study. The genomic coordinate data for the binding loci of these two sex-hormone receptors were from Kennedy et al. [78], who created a database collecting and containing the relevant data generated by ChIP-based high-throughput experiments in the published literature. Among Kennedy's data, only coordinate

Table 4.3: Proportion of positive SNPs for SA selection at the estrogen and androgen receptor binding loci

Loci	No. of SNPs	Proportion of positive SNPs	<i>P</i> value compared to background*
Estrogen receptor binding loci	5,195	0.79%	0.98
Estrogen receptor binding loci with 10Kb extention [#]	69,029	0.99%	0.97
Androgen receptor binding loci	657	1.07%	0.55
Androgen receptor binding loci with 10Kb extention [#]	8,354	0.95%	0.86

The proportion of positives among the SNPs located inside the sex-hormone receptor binding regions was compared to the background proportion. The SNPs showing $P < 0.01$ for sex-differential genotype distribution in the control individuals were defined as ‘positive’. The background proportion was 1.06%, and was the positive proportion among all the ~ 1.2 million SNPs used in the genome-wide scan.

[#]5Kb extension from each of the two boundaries of the receptor binding loci.

**P* values were generated by one-sided Fisher’s exact test.

data from in vivo experiments were employed, where there were 16,422 estrogen receptor binding loci and 1,793 androgen receptor binding loci included (accessed 1st June 2011).

However, no significant enrichment of signals for SA selection was observed at the estrogen or androgen receptor binding loci. Among the SNPs located in SHRBL, the proportions of SNPs with $P < 0.01$ for SA selection were not significantly higher than the background proportion, with *P* values of 0.98 and 0.55 for estrogen and androgen receptor binding loci, respectively. Because more than 91% of the receptor binding regions employed in the study were less than 1Kb in size, most of the regions did not contain any WTCCC SNPs. To obtain more SNPs, extension of 5Kb from each of the boundary of the binding region

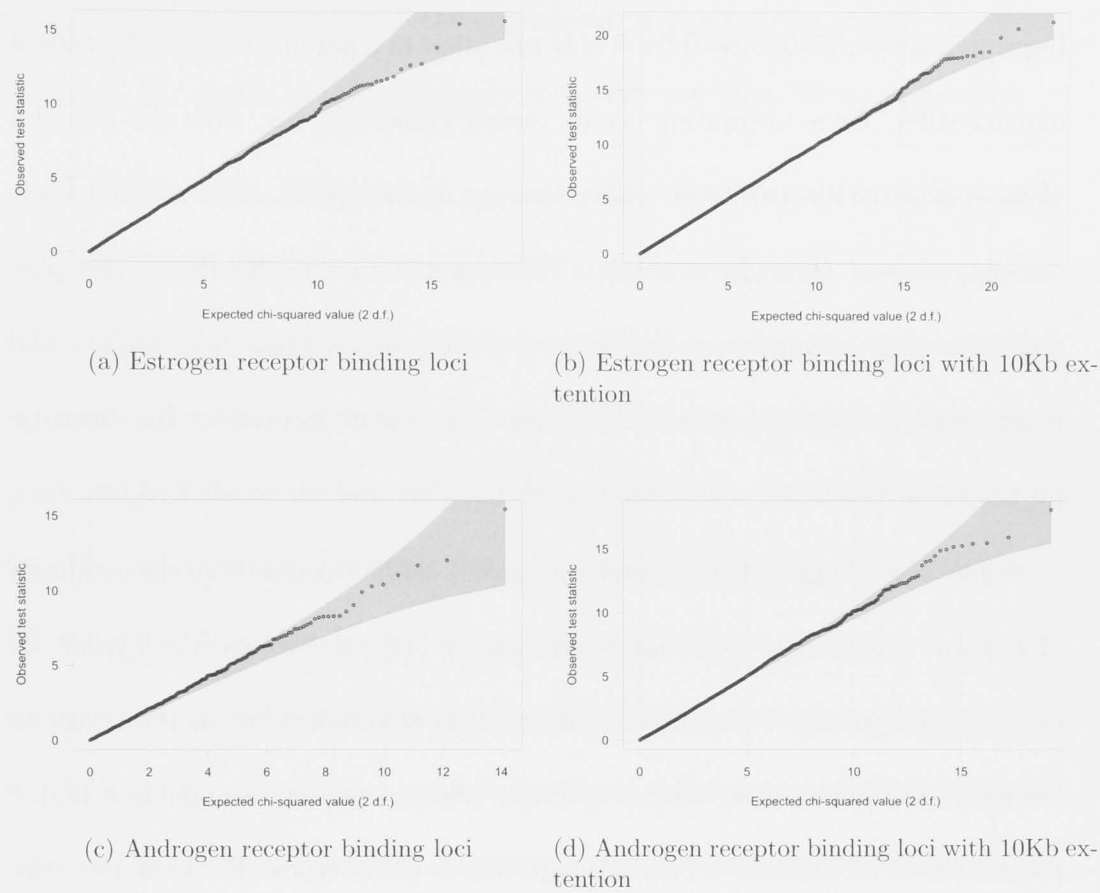


Figure 4.8: Q-Q plots for signals for SA selection from sex-hormone receptor binding loci.

was applied. However, there was still no significant enrichment observed in the extended regions. The details for the signals in SHRBL are shown in Table 4.3. Shown in the corresponding Q-Q plot, there were no obvious departures from the 95% concentration band, also indicating no significant enrichment (Figure 4.8).

4.4 No enrichment in imprinted genes

Imprinted genes are expressed from only one allele in a parent-of-origin dependent manner [108]. Gene imprinting occurs during gametogenesis, with one parental allele epigenetically marked in either the egg or the sperm. The marked allele remains silenced throughout somatic development in the offspring. The phenomenon of gene imprinting has been observed across mammals, insects and plants [54, 84, 175]. In humans, the disruption to gene imprinting has been reported to be associated with a number of disorders and cancers [25, 53].

It has been suggested that intralocus sexual conflict could drive the evolution of genomic imprinting. Day and Bonduriansky [41] put forward the hypothesis that, when a variant is experiencing differential selection between the sexes, silencing of the alleles from the less-strongly selected parent can enable a higher proportion of individuals expressing high-fitness phenotypes. For example, suppose there is an allele A which is strongly beneficial to females but neutral or a little harmful to males. In that case, Day and Bonduriansky [41] suggested that silencing of paternally inherited alleles can evolve, because daughters will benefit from exclusively expressing the maternally inherited allele (as the frequency of allele A is higher in maternal gametes than in paternal gametes), whereas the fitness of sons is roughly independent of the alleles carried at this locus.

To examine whether the imprinted genes were enriched of signals for SA selection, 55 imprinted genes of humans were collected. These genes were from the “genomic imprinting website” (<http://www.geneimprint.com/>, accessed 1st

March 2011), with status assigned as “imprinted”, and included in the Ensembl database release 59 from which the corresponding genomic coordinate data were obtained. For details of these 55 imprinted genes, see Table A3 in Appendices.

Located in these 55 imprinted genes are 2,079 WTCCC SNPs tested during the genome-wide scan. Among these SNPs, the proportion of SNPs with $P < 0.01$ for SA selection was 0.67%, which was smaller than the 1.06% of background proportion. Two-sided Fisher’s exact test for comparison of these two proportions generated a P value of 0.086. The corresponding Q-Q plot also showed part of the observed statistics were smaller than the expected values under null condition (Figure 4.9). As at the imprinted loci, only one allele is expressed, if there is selection it could be between alleles rather than between genotypes. Thus whether the SNPs in imprinted genes exhibit significant difference in allelic frequencies between the sexes was examined. However, the test statistics are not greater than their expected chi-squared values (Figure 4.10). All results indicate no evidence for signal enrichment in the regions of imprinted genes. Both test statistics shown in Q-Q plots 4.9 and 4.10 tend to be smaller than the expected chi-squared values, probably reflecting the considerable LD among the SNPs tested. Due to LD, the effective number of independent tests is smaller than the number of SNPs tested, while the expected chi-squared values are generated based on the number of SNPs tested [6].

There might be four potential reasons for no signal enrichment observed in imprinted genes. Firstly, SA selection does not result in genomic imprinting.

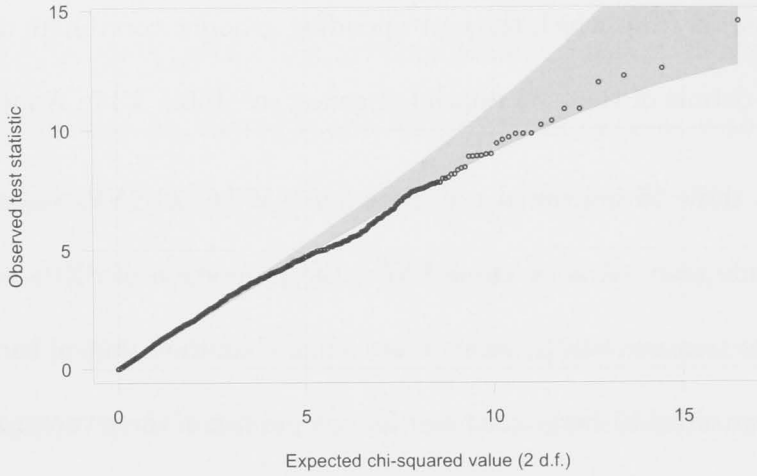


Figure 4.9: Q-Q plot for the 2,079 SNPs located inside the 55 imprinted genes. Test is for the differences in genotype frequencies between the sexes.

Secondly, only a small proportion of imprinted genes are evolutionally driven by SA selection, leading to too low statistical power for detection. In other words, more than one evolutionary mechanisms contributes to genomic imprinting [175]. Thirdly, for the imprinted genes, natural selection might be acting dominantly based on fertility fitness, while the data used in this study are only about viability fitness. Lastly, the establishment of imprinted genes only represents the past sexual antagonism or the current sexual antagonism that have been mostly resolved [39], whereas the signals detected from this study are for ongoing SA selection.

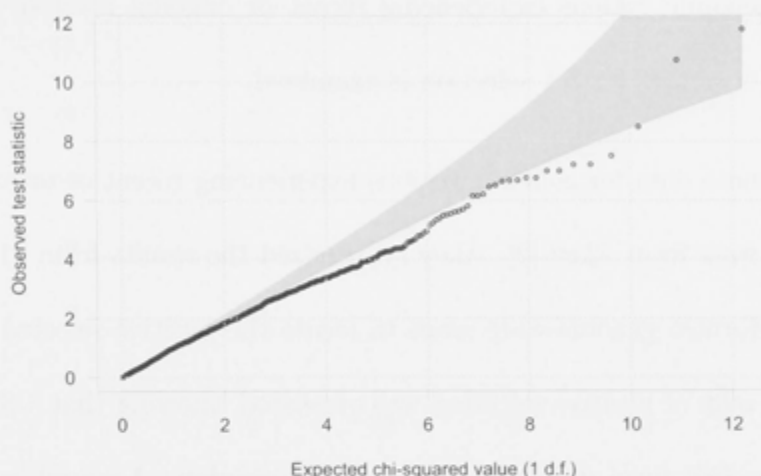


Figure 4.10: Q-Q plot for the 2,079 SNPs located inside the 55 imprinted genes. Test is for the differences in allelic frequencies between the sexes.

4.5 No enrichment in regions of positively selected selection

Whether intralocus sexual conflict is an impediment or a driver to adaptive evolution is still under debate [18]. Many researchers support the theory that sexual antagonism obstructs sexual coevolution, because simply by definition, sexual antagonism displaces both sexes from their phenotypic optimum. However, theory also suggests that sexual antagonism can promote adaptive evolution and speciation [118]. Sexual antagonism acting on a trait in one sex can do so, by reducing the average fitness of the two sexes to the bottom, and ultimately allowing that trait to evolve in a new niche [89]. One of the potential examples is that, in mammals, horns and antlers that nowadays serve as anti-predator defences could

have evolved from weapons used in male–male competition [18]. In this section, whether the genomic regions experiencing recent or ongoing positive selection were enriched of signals for SA selection is examined.

The coordinate data for genomic regions experiencing recent or ongoing positive selection were from Akey [2]. Akey synthesized the results from 21 studies, which had performed genome-wide scans to locate the positive selected regions. An integrated map of positive selection was presented, showing that $\sim 8\%$ of the human genome had been influenced by positive selection. Located in the 8% region, 92,557 WTCCC SNPs ($\sim 7.6\%$ of the total) had been tested during the genome-wide scan.

However, no significant enrichment of signals for SA selection was showed in the positively selected regions. Among the 92,557 SNPs, the proportion showing $P < 0.01$ for sex-differential genotype distributions was 1.12%, which was slightly greater than the 1.06% of background proportion. But comparison of the whole distributions of signals from the positively selected regions and from all the SNPs used in genome-wide scan generated $P=0.78$ with one-sided Kolmogorov–Smirnov test, suggesting no significant signal enrichment in the regions of positively selected selection. The Q-Q plot also showed the vast majority of the comparison results between the observed test statistics and the expected chi-squared values were located inside the 95% concentration band, with the exception that the observed statistics of ~ 10 and ~ 14 (corresponding to $P=0.006$ and $P=0.001$, 2 d.f., respectively) showed a little greater enrichment than the expected distribution,

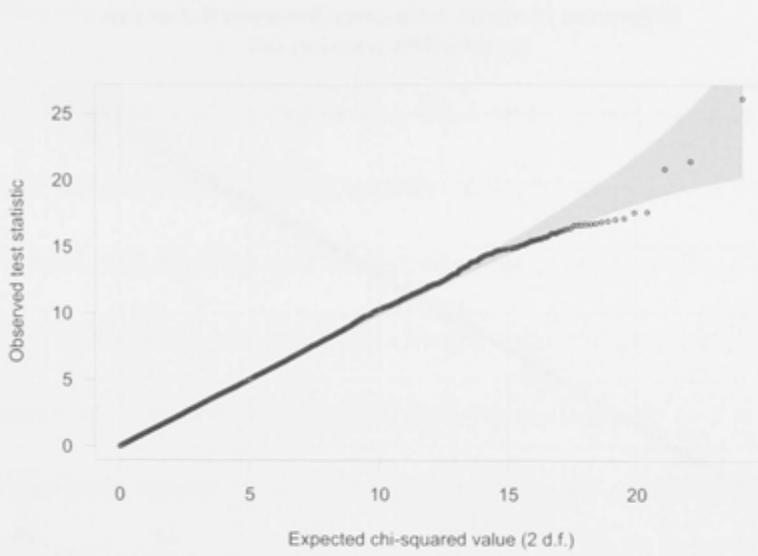


Figure 4.11: Q-Q plot for the 92,557 SNPs located in the positively selected genomic region.

and the exception that some of the observed statistics at the extreme tail showed less enrichment than the expected distribution (Figure 4.11).

As in the discussion of imprinted genes, the selection on these genomic regions might only represent the past evolutionary events, leading to no significant enrichment observed in this study. Another potential reason is, as shown in Section 4.1, the observation of signals for SA selection from this study was due to the variations causing diseases with sex-differential effect, and such variations might be under purifying selection rather than positive selection [16].

4.6 The allelic effect underlying SA selection

Natural selection is one of the forces that change allelic frequencies, driving human evolution [70]. However, whether sex-differential natural selection changes

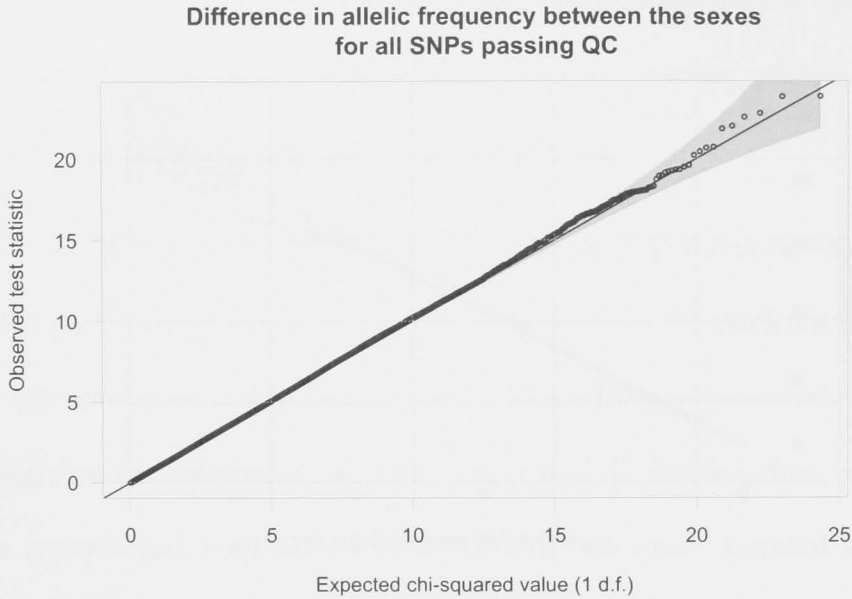


Figure 4.12: Q-Q plot for the comparison between the test statistics for differences in allele frequency between the sexes (τM_A) and the expected values. The shaded region is the 95% concentration band, and the diagonal line is $y = x$.

the allelic frequencies depends on the net effect of selection between males and females. If natural selection acts on a variant in the same direction between the sexes, despite differential magnitudes, the allelic frequency will be changed directionally, and the favored allele can be fixed eventually. However, in the case of opposite direction of natural selections, allelic frequencies may be kept stable, maintaining the genetic polymorphism in humans [39].

Because the population samples used in this study are cross-sectional, it is not possible to determine whether there are changes of allelic frequencies over generations due to SA selection. In this section, the main purpose is to study whether there is significant sex differences in allelic frequencies in humans, and whether there were allelic effects that significantly contributed to the sex-differential geno-

type distribution observed from the study.

The allelic frequencies between the sexes were compared simultaneously at all the ~ 1.2 million SNPs passing quality control process. The statistical tests were performed with Huttley and Wilson's log-linear models described in Chapter 3.1.1. If the individuals are randomly sampled from a population, in which no allelic frequencies between the sexes are presented across the whole genome, then the statistical test statistics are expected to be in the chi-squared distribution. The comparison between the observed test statistics and the expected chi-squared values is showed in Figure 4.12. This figure shows that the observed statistics greater than 14 slightly departed over the diagonal line (where $y = x$). However, the most majority were still located in the 95% concentration band, with slight departure for the values from ~ 14 to 17 (approximately corresponding to P values from $2E-4$ to $4E-5$, 1 d.f.). These results suggest that there might be differential allelic frequencies between the sexes caused by natural selection, with low confidence.

Other evidence for the existence of sex-differential allele frequencies is from the comparison between the statistics of unequal allele frequencies (τM_A) and Hardy-Weinberg (HW) disequilibria (τM_{AA}). In Huttley and Wilson's log-linear models, the deviance measuring genotypic discordance between the sexes is the sum of τM_A and τM_{AA} . In addition, the distributions of statistics of τM_A and τM_{AA} are expected to be the same, if they contribute equally to the genotypic difference between the sexes. However, the comparison showed that the statistics

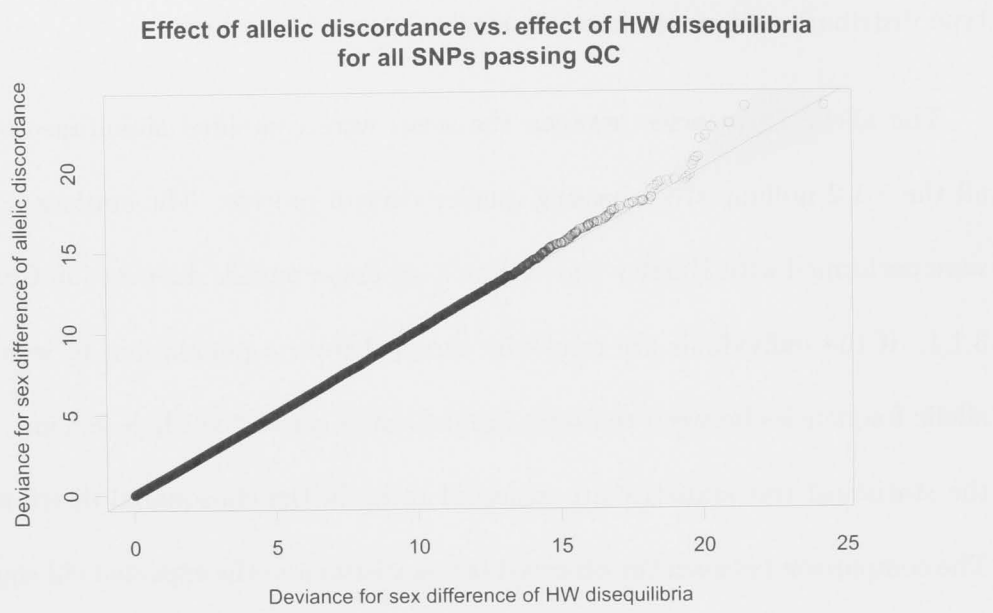
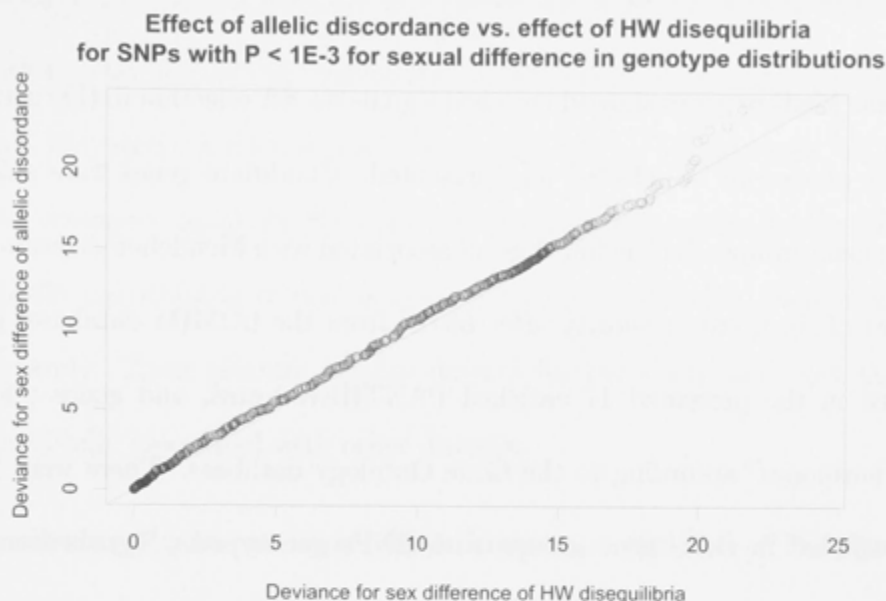
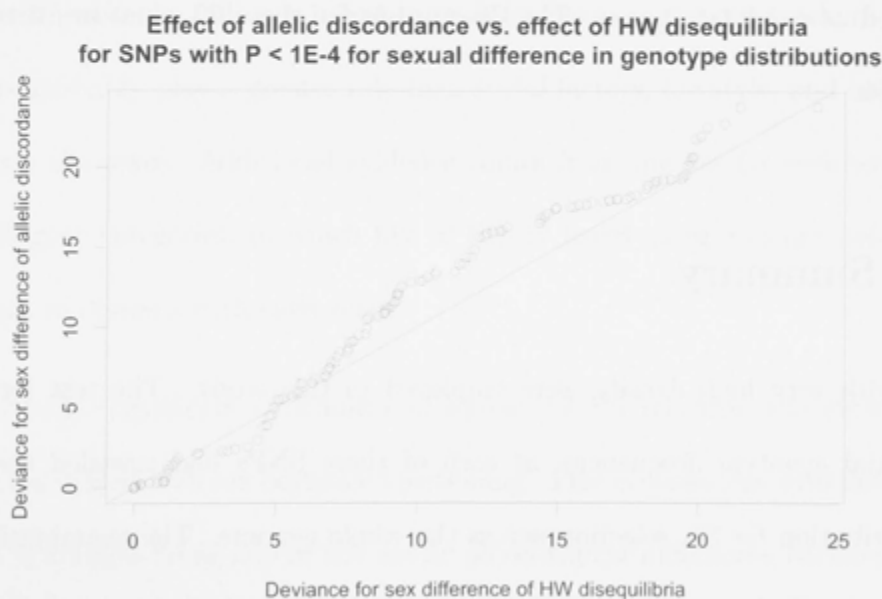


Figure 4.13: Q-Q plot for the comparisons between values of τM_A (Y axis) and of τM_{AA} (X axis), for all the SNPs passing the QQ and used in the genome-wide scan. The diagonal line is $y = x$.

of τM_A were greater than the statistics of τM_{AA} in the tail of the Q-Q plot (Figure 4.13). The statistics located in the tail presented the SNPs potentially experiencing the strongest SA selection. Consistently, τM_A had obviously much greater values than the τM_{AA} did, when only the SNPs exhibiting $P < 1\text{E-}3$ and $P < 1\text{E-}4$ for SA selection were included (Figure 4.14). These results suggest the existence of sex-differential allele frequencies and that they contribute to the sex-differential genotype distribution observed in the study.



(a)



(b)

Figure 4.14: Q-Q plot for the comparisons between values of τM_A (Y axis) and of τM_{AA} (X axis), only for the SNPs with P values less than $1E-3$ (a) and less than $1E-4$ (b) for the potential ongoing SA selection. The diagonal line is $y = x$.

4.7 Candidate genes

Since some kinds of genes showed enriched signals for SA selection in this chapter, candidate genes can be selected and presented. Candidate genes were selected from the gene groups, that included genes associated with Mendelian diseases with the onset of “in utero or shortly after birth” from the hOMIM database, genes embodied in the presented 17 enriched PANTHER terms, and genes relevant to “sex hormones” according to the Gene Ontology database. There were 1,433 genes included in these gene groups with SNPs genotyped. Signals from the SNPs in these genes were adjusted by permutation test (details for this method see Chapter 3.1.2.3), and 22 genes with adjusted P value less than 0.01 were presented as candidate genes. The description for these 22 genes are listed in Table 4.4.

4.8 Summary

SNPs with very high density were employed in this study. The test for sex-differential genotype frequencies at each of these SNPs had revealed the signal distribution for SA selection across the whole genome. The examination of whether specified genes or genomic regions were enriched of signal is helpful in exploring the characteristics of variants experiencing SA selection, and in testing several evolutionary hypotheses.

Firstly, the characteristics of variants experiencing SA selection revealed in

this study are consistent with the hypothesis that it is sex difference of viability or mortality leading to the observation of differences in genotype distributions between the sexes. Evidence has been presented that the genomic regions exhibiting strongest signals for SA selection during the genome-wide scan were sex-specifically contributing to one or more of the seven common diseases involved in the study. These genomic regions deserve further study, and they might be sex-specifically associated with other diseases.

Furthermore, genes associated with diseases are enriched of the signals for SA selection. For the SNPs located in the Mendelian disease genes with early onset, there was significant enrichment of signals for SA selection. This result is consistent with the different infant mortalities between the sexes, where genetic factors probably play a greater role than social factors, life style, and behaviour between the sexes. Additional evidence comes from the list for enriched PANTHER gene categories, in which five of the 17 listed categories are potentially relevant to diseases with early onset.

Secondly, significant enrichment of signals for SA selection was shown from genes associated with sex hormone functioning. This is consistent with differences in sex hormones being one of the major physiological differences between males and females, and potentially contributing to the sex-differential viabilities and mortalities. In turn, the observation of signal enrichment from genes associated with sex hormones and diseases made the results of this study more convincing. No variations from these genes were identified in the genome-wide scan, reflecting

potentially real signals with small effect were hidden below the stringent threshold for genome-wide scan [66]. The candidate genes from these enriched gene categories are presented.

Thirdly, three evolutionary hypotheses have been tested. However, no positive results were obtained for the examinations for enrichment of signals in cis-regulatory elements of sex-hormone pathway, imprinted genes, or in the positively selected genomic regions. Perhaps these were due to the past evolutionary events that were hardly be replicated with the data of ongoing activities employed in the study.

Lastly, this study showed that due to the sexual antagonism, the differences, not only in genotype frequencies, but also in allelic frequencies, can be observed in human random samples. This discovery has important implication, especially for those studies that are sensitive to the ratio of the sexes.

Table 4.4: Candidates genes selected from the gene groups enriched of signals for SA selection. For details for the selection see the context.

Gene Name	Reported OMIM disease (Early onset?)	Knockout mouse phenotype	Enriched PANTHER terms	GO term of sex- hormone
MAD1L1 (mitotic arrest-deficient 1, yeast, homolog-like 1)	Prostate cancer (Not early onset)	Lethality/prenatal- perinatal; tumorigenesis	Mitosis	-
NKX3-1 (NK3 homeobox 1; NKX3-1)	-	Lethality/prenatal- perinatal; tumorigenesis	Heart development	-
CDC37L1 (CDC37-like 1)	-	-	Other signal transduction	-
NANOS1 (NANOS, Drosophila, homolog of, 1)	-	No phenotypic alterations	Anterior/posterior patterning	-
SLC35B4 (solute carrier family 35 (UDP- xylose/UDP-n- acetylglucosamine transporter), member B4)	-	-	Carbohydrate transport	-
SLC22A25 (organic anion transporter UST6)	-	-	Anion transport	-

Gene Name	Reported OMIM disease (Early onset?)	Knockout mouse phenotype	Enriched PANTHER terms	GO term of sex-hormone
PRKAG2 (protein kinase, amp-activated, noncatalytic, gamma-2)	Cardiomyopathy, familial hypertrophic, 6; glycogen storage disease of heart, lethal congenital (Early onset)	-	-	-
TGFB2 (transforming growth factor, beta-2)	-	lethality/prenatal-perinatal;	-	Response to estrogen stimulus
ADSL (adenylosuccinate lyase)	Adenylosuccinase deficiency (Early onset)	-	-	-
ARSA (arylsulfatase A)	Metachromatic leukodystrophy (Early onset)	Affected Systems: nervous system, behavior, homeostasis, hearing/vestibular/ear	Sulfur metabolism	-
KIF15 (kinesin family member 15)	-	-	Mitosis	-
CHST11 (carbohydrate sulfotransferase 11)	-	Lethality/prenatal perinatal	Sulfur metabolism	-
PROX2 (prospero homeobox 2)	-	No phenotypic alterations	Gut mesoderm development	-

Gene Name	Reported OMIM disease (Early onset?)	Knockout mouse phenotype	Enriched PANTHER terms	GO term of sex- hormone
BCL2L10 (BCL2-like 10)	-	No phenotypic alterations	Other signal transduction	-
CEP135 (centrosomal protein, 135-KD)	-	-	Mitosis	-
AGPS (alkylglycerone- phosphate synthase)	Rhizomelic chondrodyspla- sia punctata, type 3 (Early onset)	-	Other lipid, fatty acid and steroid metabolism; Electron transport	-
PDGFB (platelet-derived growth factor, beta polypeptide)	Dermatofibrosarcoma protuberans (Not early onset)	Lethality/prenatal- perinatal	-	Response to estrogen stimulus
IL6ST (interleukin 6 signal transducer)	-	Lethality/prenatal- perinatal	MAPKKK cascade	-
CHST13 (carbohydrate sulfotransferase 13)	-	-	Sulfur metabolism	-
SLC22A9 (solute carrier family 22 (organic anion/cation transporter), member 9)	-	-	Anion transport	-

Gene Name	Reported OMIM disease (Early onset?)	Knockout mouse phenotype	Enriched PANTHER terms	GO term of sex- hormone
SLC22A10 (solute carrier family 22 (organic anion/cation transporter), member 10)	-	-	Anion transport	-
UROCI (urocanase domain- containing protein 1)	Urocanase deficiency (Early onset)	-	-	-

Chapter 5

Conclusions

The two sexes differ in many ways. In such two distinct environments, a genetic variant shared by the two sexes may function differentially, resulting in different fitnesses in the two sexes. Sexual antagonism occurs when natural selection on a shared trait in one sex displaces the other sex from its optimum. There has been substantial evidence for SA selection from a broad range of taxa. But this has not been widely studied in humans.

This study aimed to detect ongoing SA selection at the viability level in humans. A variant experiencing such sexual antagonism could result in differential or opposite effects on viability or mortality between males and females. The main material in the study was the autosomal SNP data for the control individuals from the WTCCC, which are representative of a random sample of white Europeans. The discordance in genotype distribution between the sexes was used as an indicator for SA selection. In this study, the genotype distribution

between the sexes at ~ 1.2 million SNPs across the genome has been compared, and the annotation for genes and genomic regions has also been integrated to interpret the genetics underlying SA selection.

5.1 Discoveries from the study

Following the examination for the signals around the two SNPs, which reportedly exhibited significant difference in genotype distributions between the sexes in general human populations, BRCA2 was presented as a candidate gene experiencing SA selection. It has been reported that BRCA2 is a tumour-suppressor gene, and deleterious mutations in BRCA2 are the principal cause of hereditary breast and/or ovarian cancer. It is likely that the sex-differential effects of variants in BRCA2 caused the sexual antagonism observed in the study.

Additional direct evidence for SA selection in humans is from the genome-wide scan. The comparisons of genotype distributions between the sexes at the ~ 1.2 million SNPs across the whole genome showed significantly more SNPs of sexual discordance than expected. Eight independent genomic regions with the strongest signals, as well as the candidate genes from these regions, were presented in the study. Evidence of allele frequency variation between the sexes in control samples due to SA selection was also presented.

The inspection of signals from disease genes with early onset and from genes associated with sex hormone functioning showed enrichment of signals for SA selection. The former supports the hypothesis that it is the sex difference in viability

or mortality leading to the observation of differences in genotype distributions between the sexes. The latter is consistent with differences in sex hormones being one of the major differences between males and females, and these differences potentially contributing to the sex-differential viability and mortality. In turn, these results provide support for the existence of SA selection, making the findings from this study more reliable. Candidate genes from these enriched genomic regions and from other enriched gene categories embodied by the PANTHER system are presented in the study.

Lastly, the examination for SNPs from the specific genomic regions, including cis-regulatory elements of sex-hormone pathway, imprinted genes, and the positively selected regions, shows no trace of ongoing SA selection.

5.2 Importance of the discoveries

The evidence presented in this study shows that we humans, just like other sexually reproducing organisms, are experiencing SA selection. It also supports the hypothesis that SA selection is inevitable and ubiquitous due to the differences between males and females. Natural selection of such pattern might have been acting on humans since the origin of sexual reproduction. Furthermore, given the sample size ($\sim 4,800$ individuals) used in the study, the ongoing SA selection is still detectable, reflecting the potentially great impact on human evolution.

The presentation of candidate genes experiencing SA selection revealed in this study is significant. Although substantial evidence for the existence of SA selec-

tion has been presented, the vast majority of previous studies only showed the sex-differential phenotypic effects of intralocus sexual conflict, without exploring the genetic basis, let alone the localization of the genes involved. Thus the presented candidate genes from this study are important materials for studying the genetic mechanisms underlying SA selection, and for studying the evolution of genes experiencing SA selection, at the viability level.

Diseases were shown to be underlying SA selection. This has important implications to human health. Among the candidate genes presented in the study, the majority were not yet reported to be associated with known diseases. They thus are candidates responsible for the unknown diseases with sex-specific effects. In particular, the candidate genes provide clues for exploring the genetic basis contributing to the different infant mortalities between the sexes.

Novel quality control filters, to exclude regions of genomic duplication, were utilized in the study. These filters are effective in screening the duplicon SNPs, which can not be genotyped correctly. These novel QC measures can also be applied to other genome-wide studies especially where the two compared groups have marked differences in the ratio of the two sexes.

This study also provides a fundamental understanding of allele and genotype frequency distributions in human populations. Even in the control or random samples, there is still the inherent variation of allele and genotype frequencies between the sexes. This has important implication, for genetic studies that ignore such sex-specific effects in their design and interpretation could lead to spurious

outcomes.

5.3 Further study

The findings regarding SA selection from this study need to be replicated in independent samples in the future. Bigger sample size might be needed, given the modest effect of potential SA selection. With candidate genomic regions and genes presented in the study, fine-mapping (e.g. sequencing the whole region) will be needed to determine the causative variants. Lastly, because this study only identifies natural selection for viability, whereas SA selection arising from differential fertility might be stronger [39], more comprehensive studies in humans are necessary in the future.

Appendices

Table A1: Positive ($P < 0.01$) proportions for SNPs from each of the 242 PANTHER biological process categories (version 6.1), in the ascending order of q values.

The q values were the multiple-testing correction for the total 242 categories, for comparisons of positive proportions between the specified category and the background. The background proportion was 1.06%, and was the proportion of SNPs with $P < 0.01$ among all the SNPs used in the genome-wide scan.

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Steroid hormone metabolism	36	622	6.91	3.89E-18
Electron transport	140	3478	2.73	3.05E-13
mRNA polyadenylation	25	433	6.46	2.35E-11
Other lipid, fatty acid and steroid metabolism	25	684	4.53	5.53E-09
Mitosis	138	3673	2.28	2.26E-08
MAPKKK cascade	170	6058	1.91	2.90E-07
Anion transport	78	2782	2.19	1.30E-05
Sulfur metabolism	49	2262	2.21	9.52E-05
Other signal transduction	64	1416	2.47	0.00025
Fatty acid beta-oxidation	23	874	2.51	0.00742
Heart development	48	1863	1.98	0.00817
Gametogenesis	26	605	2.80	0.00817
Carbohydrate transport	48	1602	2.05	0.00817
Phosphate transport	16	335	3.28	0.02261
Gut mesoderm development	14	457	2.84	0.02822
Anterior/posterior patterning	58	2057	1.79	0.03281
Receptor protein serine/threonine kinase signaling pathway	39	2192	1.73	0.04930
Amino acid activation	38	755	2.25	0.05611
Other sensory perception	14	1089	2.02	0.05631

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Amino acid biosynthesis	41	965	2.07	0.05709
Vitamin biosynthesis	17	519	2.50	0.05709
Lipid metabolism	143	5866	1.43	0.06155
Induction of apoptosis	149	3956	1.51	0.06155
Other immune and defense	145	3369	1.54	0.07153
Non-vertebrate process	18	207	3.38	0.07842
Other sulfur metabolism	16	558	2.32	0.08080
RNA localization	32	981	1.93	0.10332
Extracellular transport and import	79	3237	1.51	0.10352
Skeletal development	115	3808	1.47	0.10754
Asymmetric protein localization	19	927	1.94	0.10754
Carbohydrate metabolism	136	3267	1.49	0.10836
Vitamin metabolism	19	289	2.76	0.10930
Other neuronal activity	126	9313	1.30	0.11340
Extracellular matrix protein-mediated signaling	61	3464	1.47	0.11832
Polyphosphate catabolism	3	91	4.39	0.12732
Other protein targeting and localization	18	253	2.76	0.14127
Pterin metabolism	8	97	4.12	0.14127
Developmental processes	469	18452	1.22	0.14127
Neurogenesis	549	31783	1.18	0.14127
DNA replication	120	2186	1.50	0.20798
Protein phosphorylation	603	28648	1.17	0.21180
Muscle development	128	3419	1.40	0.22278
Complement-mediated immunity	52	1131	1.67	0.22389
Proteolysis	882	21483	1.19	0.23601
Vitamin catabolism	1	30	6.66	0.24043
Other pathways of electron transport	19	302	2.31	0.24844
Small molecule transport	127	4171	1.34	0.27642
Vision	178	5332	1.29	0.32637
Metabolism of cyclic nucleotides	43	3563	1.34	0.32637
Inhibition of apoptosis	120	2988	1.37	0.32860
Hearing	29	996	1.60	0.35498
Mesoderm development	134	6025	1.26	0.38724
JAK-STAT cascade	78	1569	1.46	0.38791
Embryogenesis	130	4960	1.27	0.43038
mRNA capping	3	104	2.88	0.45872
Fatty acid metabolism	107	2374	1.34	0.48685
JNK cascade	61	1469	1.42	0.48719

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Miscellaneous	129	3075	1.30	0.50611
General mRNA transcription activities	43	533	1.68	0.50611
Hematopoiesis	65	1663	1.38	0.52919
Angiogenesis	50	2191	1.32	0.57538
rRNA metabolism	63	724	1.51	0.61924
Apoptotic processes	17	268	1.86	0.62829
Porphyrin metabolism	13	209	1.91	0.70564
Biological process unclassified	8809	164606	1.08	0.70564
Oncogene	96	3358	1.22	0.78105
Cell motility	322	15336	1.12	0.85744
Stress response	186	4880	1.16	0.92620
Phospholipid metabolism	134	4614	1.17	0.92620
Nerve-nerve synaptic transmission	72	5174	1.15	0.94682
Oxidative phosphorylation	69	669	1.34	0.98649
Other oncogenesis	47	1193	1.25	0.99837
Other intracellular signaling cascade	206	10901	1.11	1
Regulation of carbohydrate metabolism	12	513	1.36	1
Protein modification	196	5630	1.13	1
Protein methylation	10	186	1.61	1
Coenzyme metabolism	53	1304	1.22	1
Constitutive exocytosis	11	187	1.60	1
Sensory perception	23	721	1.24	1
Purine metabolism	61	1627	1.16	1
Cytokinesis	103	1451	1.17	1
Regulated exocytosis	46	2188	1.14	1
Oncogenesis	154	6330	1.10	1
Receptor mediated endocytosis	102	3348	1.10	1
Neurotransmitter release	93	6934	1.08	1
Cell structure	619	21443	1.07	1
General vesicle transport	227	6857	1.07	1
Fatty acid biosynthesis	14	517	1.16	1
Cell communication	309	16318	1.07	1
Steroid hormone-mediated signaling	44	984	1.11	1
Other receptor mediated signaling pathway	197	5665	1.07	1
Cell adhesion	531	38805	1.06	1
Pre-mRNA processing	36	350	1.14	1

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Chromatin packaging and remodeling	202	3180	1.06	1
Detoxification	67	1872	1.06	1
Other intracellular protein traffic	59	2159	1.06	1
Amino acid transport	42	1698	1.06	1
Macrophage-mediated immunity	136	1894	1.05	1
Cell structure and motility	183	7673	1.05	1
MHCI-mediated immunity	18	172	1.16	1
Other carbohydrate metabolism	54	1618	1.05	1
Pentose-phosphate shunt	8	78	1.28	1
Other transport	57	1155	1.03	1
Polyphosphate biosynthesis	3	80	1.25	1
Apoptosis	193	5743	1.04	1
DNA metabolism	30	482	1.03	1
Other mRNA transcription	19	796	1.00	1
Glycogen metabolism	43	797	1.00	1
Protein targeting	101	3517	1.02	1
Nucleoside, nucleotide and nucleic acid transport	42	805	0.99	1
Gluconeogenesis	17	607	0.98	1
mRNA transcription regulation	1303	29784	1.04	1
Coenzyme and prosthetic group metabolism	10	518	0.96	1
Glucose homeostasis	15	723	0.96	1
Other cell cycle process	9	213	0.93	1
Other apoptosis	12	213	0.93	1
Other nucleoside, nucleotide and nucleic acid metabolism	28	1037	0.96	1
Cell cycle control	380	10926	1.02	1
Immunity and defense	275	6935	1.00	1
Glycolysis	41	546	0.91	1
Signal transduction	272	11092	1.01	1
Endoderm development	11	338	0.88	1
Tricarboxylic acid pathway	31	760	0.92	1
Other metabolism	433	9305	1.01	1
mRNA transcription	253	5659	0.98	1
Segment specification	92	1917	0.93	1
Sex determination	9	126	0.79	1
Protein-lipid modification	21	480	0.83	1
DNA recombination	38	1126	0.88	1
Protein biosynthesis	464	3360	0.95	1

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Regulation of vasoconstriction, dilation	40	816	0.85	1
mRNA splicing	188	3674	0.95	1
Other amino acid metabolism	28	607	0.82	1
Cytokine and chemokine mediated signaling pathway	225	4088	0.95	1
NO mediated signal transduction	7	413	0.72	1
Peroxisome transport	14	302	0.66	1
Reverse transcription	8	305	0.65	1
Intracellular signaling cascade	235	8253	0.95	1
Steroid metabolism	59	1456	0.82	1
Amino acid catabolism	43	1130	0.79	1
Other coenzyme and prosthetic group metabolism	9	185	0.54	1
Cell adhesion-mediated signaling	321	23560	0.99	1
Nucleoside, nucleotide and nucleic acid metabolism	244	4262	0.89	1
Other carbon metabolism	75	1737	0.80	1
Transport	452	11618	0.95	1
NF-kappaB cascade	71	1199	0.75	1
Muscle contraction	181	9892	0.94	1
Lysosome transport	11	214	0.46	1
Protein metabolism and modification	68	1799	0.77	1
Interferon-mediated immunity	62	645	0.62	1
Other homeostasis activities	61	1258	0.71	1
Calcium mediated signaling	107	8018	0.91	1
tRNA metabolism	41	397	0.50	1
Cholesterol metabolism	64	1976	0.75	1
Tumor suppressor	75	2750	0.8	1
Protein targeting and localization	84	2538	0.78	1
Cation transport	443	22691	0.96	1
Endocytosis	112	3545	0.81	1
Other blood circulation and gas exchange activity	16	267	0.37	1
Oogenesis	62	3890	0.82	1
Blood clotting	86	2370	0.75	1
Other developmental process	94	3912	0.81	1
B-cell- and antibody-mediated immunity	86	1699	0.70	1
Vitamin/cofactor transport	42	604	0.49	1

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Homeostasis	59	2421	0.74	1
Nitric oxide biosynthesis	6	305	0.32	1
Olfaction	195	510	0.39	1
Action potential propagation	20	522	0.38	1
Regulation of lipid, fatty acid and steroid metabolism	26	679	0.44	1
Synaptic transmission	74	8228	0.85	1
G-protein mediated signaling	786	18016	0.91	1
Translational regulation	74	1111	0.54	1
Lipid, fatty acid and steroid metabolism	37	1116	0.53	1
Acyl-CoA metabolism	19	365	0.27	1
Ion transport	43	1262	0.55	1
Cytokine/chemokine mediated immunity	113	1393	0.57	1
Amino acid metabolism	57	1003	0.49	1
Protein complex assembly	60	865	0.46	1
Lipid and fatty acid transport	120	3705	0.72	1
T-cell mediated immunity	128	2549	0.66	1
Cell proliferation and differentiation	926	25159	0.92	1
Mitochondrial transport	24	609	0.32	1
Fertilization	35	774	0.38	1
DNA repair	148	3781	0.71	1
Natural killer cell mediated immunity	69	634	0.31	1
Cell surface receptor mediated signal transduction	158	6422	0.77	1
Sulfur redox metabolism	19	456	0.21	1
Exocytosis	92	2456	0.61	1
Meiosis	72	1571	0.50	1
RNA catabolism	61	1015	0.39	1
Protein disulfide-isomerase reaction	17	512	0.19	1
Other protein metabolism	43	746	0.26	1
Nuclear transport	121	1394	0.43	1
Neuromuscular synaptic transmission	22	1397	0.42	1
Protein acetylation	27	548	0.18	1
Other polysaccharide metabolism	123	5424	0.70	1
Pain sensation	29	566	0.17	1
Neuronal activities	127	6369	0.72	1

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Granulocyte-mediated immunity	57	585	0.17	1
Cell cycle	112	3329	0.60	1
Ectoderm development	94	1929	0.46	1
Monosaccharide metabolism	37	1010	0.29	1
Protein folding	167	1721	0.40	1
Spermatogenesis and motility	106	2587	0.50	1
Phosphate metabolism	80	3081	0.51	1
mRNA transcription initiation	59	1388	0.28	1
Calcium ion homeostasis	34	1430	0.27	1
Ligand-mediated signaling	380	10153	0.71	1
Receptor protein tyrosine kinase signaling pathway	190	11344	0.69	1
Phagocytosis	38	2876	0.38	1
Protein glycosylation	178	7654	0.61	1
Intracellular protein traffic	183	5066	0.49	1
Pyrimidine metabolism	40	1179	0.08	1
Determination of dorsal/ventral axis	14	1382	0.07	1
Chromosome segregation	105	1527	0.06	1
Fatty acid desaturation	4	49	0.0	1
Ferredoxin metabolism	4	53	0	1
Antioxidation and free radical removal	35	470	0	1
Other nitrogen metabolism	2	10	0	1
MHCII-mediated immunity	24	427	0	1
Growth factor homeostasis	7	51	0	1
mRNA end-processing and stability	16	267	0	1
Regulation of nucleoside, nucleotide metabolism	13	511	0	1
Chemosensory perception	2	11	0	1
Nitrogen utilization	4	7	0	1
Pinocytosis	5	44	0	1
Blood circulation and gas exchange	28	411	0	1
Other steroid metabolism	12	110	0	1
mRNA transcription termination	9	57	0	1
Protein ADP-ribosylation	10	187	0	1
Other phosphate metabolism	1	12	0	1
Taste	7	71	0	1
Regulation of phosphate metabolism	3	21	0	1
DNA degradation	10	108	0	1

Biological process	No. of genes	No. of SNPs included	Proportion of positive SNPs (%)	Q value
Lipid and fatty acid binding	24	217	0	1
Bile acid metabolism	1	3	0	1
Nitrogen metabolism	5	83	0	1
Pheromone response	7	22	0	1
Amino acid metabolism regulation	2	13	0	1
mRNA transcription elongation	31	323	0	1
Disaccharide metabolism	5	87	0	1
Lactation, mammary development	12	272	0	1

Table A2: The 14 terms of biological process from the Gene Ontology regarding sex hormone

Term	Description
Androgen metabolic process	The chemical reactions and pathways involving androgens, C19 steroid hormones that can stimulate the development of male sexual characteristics.
Androgen biosynthetic process	The chemical reactions and pathways resulting in the formation of androgens, C19 steroid hormones that can stimulate the development of male sexual characteristics.
Androgen catabolic process	The chemical reactions and pathways resulting in the breakdown of androgens, C19 steroid hormones that can stimulate the development of male sexual characteristics.
Estrogen metabolic process	The chemical reactions and pathways involving estrogens, C18 steroid hormones that can stimulate the development of female sexual characteristics. Also found in plants.
Estrogen biosynthetic process	The chemical reactions and pathways resulting in the formation of estrogens, C18 steroid hormones that can stimulate the development of female sexual characteristics. Also found in plants.
Estrogen catabolic process	The chemical reactions and pathways resulting in the breakdown of estrogens, C18 steroid hormones that can stimulate the development of female sexual characteristics. Also found in plants.
Cellular response to estrogen stimulus	A change in state or activity of a cell (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of stimulus by an estrogen, C18 steroid hormones that can stimulate the development of female sexual characteristics.

Term	Description
Response to estrogen stimulus	A change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of stimulus by an estrogen, C18 steroid hormones that can stimulate the development of female sexual characteristics.
Development of primary male sexual characteristics	The process whose specific outcome is the progression of the primary male sexual characteristics over time, from their formation to the mature structures. The primary male sexual characteristics are the testes, and they develop in response to sex hormone secretion.
Development of primary female sexual characteristics	The process whose specific outcome is the progression of the primary female sexual characteristics over time, from their formation to the mature structure. The primary female sexual characteristics are the ovaries, and they develop in response to sex hormone secretion.
Development of primary sexual characteristics	The process whose specific outcome is the progression of the primary sexual characteristics over time, from their formation to the mature structures. The primary sexual characteristics are the testes in males and the ovaries in females and they develop in response to sex hormone secretion.
Development of secondary female sexual characteristics	The process whose specific outcome is the progression of the secondary female sexual characteristics over time, from their formation to the mature structures. In female humans, these include growth of axillary and pubic hair, breast development and menstrual periods. Their development occurs in response to sex hormone secretion.
Development of secondary male sexual characteristics	The process whose specific outcome is the progression of the secondary male sexual characteristics over time, from their formation to the mature structures. In male humans, these include growth of axillary, chest, and pubic hair, voice changes, and testicular/penile enlargement. Development occurs in response to sex hormone secretion.
Development of secondary sexual characteristics	The process whose specific outcome is the progression of the secondary sexual characteristics over time, from their formation to the mature structures. In humans, these include growth of axillary, chest, and pubic hair, voice changes, testicular/penile enlargement, breast development and menstrual periods. Development occurs in response to sex hormone secretion.

The descriptions for the 14 terms were from the Gene Ontology database.

Table A3: Description for the selected imprinted genes

Ensembl ID	Gene Name	Location	Expressed Allele
ENSG00000078900	TP73	1p36.32	Maternal
ENSG00000162595	DIRAS3	1p31.3	Paternal
ENSG00000177432	NAP1L5	4q22.1	Paternal
ENSG00000118495	PLAGL1	6q24.2	Paternal
ENSG00000246840	HYMAI	6q24.2	Paternal
ENSG00000112499	SLC22A2	6q25.3	Maternal
ENSG00000146477	SLC22A3	6q25.3	Maternal
ENSG00000132437	DDC	7p12.1	Isoform Dependent
ENSG00000106070	GRB10	7p12.1	Isoform Dependent
ENSG00000105825	TFPI2	7q21.3	Maternal
ENSG00000127990	SGCE	7q21.3	Paternal
ENSG00000242265	PEG10	7q21.3	Paternal
ENSG00000158528	PPP1R9A	7q21.3	Maternal
ENSG00000105880	DLX5	7q21.3	Maternal
ENSG00000128510	CPA4	7q32.2	Maternal
ENSG00000106484	MEST	7q32.2	Paternal
ENSG00000246011	MESTIT1	7q32.2	Paternal
ENSG00000174595	KLF14	7q32.3	Maternal
ENSG00000198010	DLGAP2	8p23.3	Paternal
ENSG00000169427	KCNK9	8q24.3	Maternal
ENSG00000165029	ABCA1	9q31.1	Null
ENSG00000198825	INPP5F	10q26.11	Paternal
ENSG00000229414	KCNQ1OT1	11p15.4	Paternal
ENSG00000130600	H19	11p15.5	Maternal
ENSG00000167244	IGF2	11p15.5	Paternal
ENSG00000099869	IGF2AS	11p15.5	Paternal
ENSG00000129965	INS	11p15.5	Paternal
ENSG00000053918	KCNQ1	11p15.5	Maternal
ENSG00000237941	KCNQ1DN	11p15.4	Maternal
ENSG00000129757	CDKN1C	11p15.4	Maternal
ENSG00000110628	SLC22A18	11p15.4	Maternal
ENSG00000181649	PHLDA2	11p15.4	Maternal
ENSG00000021762	OSBPL5	11p15.4	Maternal
ENSG00000184937	WT1	11p13	Paternal
ENSG00000139194	RBP5	12p13.31	Maternal
ENSG00000214548	MEG3	14q32.2	Maternal
ENSG00000185559	DLK1	14q32.2	Paternal
ENSG00000182636	NDN	15q11.2	Paternal

Ensembl ID	Gene Name	Location	Expressed Allele
ENSG00000214265	SNURF	15q11.2	Paternal
ENSG00000238810	SNORD107	15q11.2	Paternal
ENSG00000239014	SNORD108	15q11.2	Paternal
ENSG00000179455	MKRN3	15q11.2	Paternal
ENSG00000179449	MAGEL2	15q11.2	Paternal
ENSG00000128739	SNRPN	15q11.2	Paternal
ENSG00000114062	UBE3A	15q11.2	Maternal
ENSG00000206190	ATP10A	15q12	Maternal
ENSG00000167522	ANKRD11	16q24.3	Maternal
ENSG00000183791	TCEB3C	18q21.1	Maternal
ENSG00000198300	ZIM2	19q13.43	Paternal
ENSG00000083844	ZNF264	19q13.43	Maternal
ENSG00000166619	BLCAP	20q11.23	Isoform Dependent
ENSG00000053438	NNAT	20q11.23	Paternal
ENSG00000185513	L3MBTL	20q13.12	Paternal
ENSG00000235590	GNASAS	20q13.32	Paternal
ENSG00000087460	GNAS	20q13.32	Isoform Dependent

These genes were from the genomic imprinting website
(<http://www.geneimprint.com/>, accessed 1st March 2011).

Bibliography

- [1] 1000 Genomes Project Consortium, Richard M Durbin, Gonçalo R Abecasis, David L Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010.
- [2] Joshua M Akey. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res*, 19(5):711–22, May 2009.
- [3] Yannis Alamanos and Alexandros A Drosos. Epidemiology of adult rheumatoid arthritis. *Autoimmun Rev*, 4(3):130–6, Mar 2005.
- [4] Andre Aleman, René S Kahn, and Jean-Paul Selten. Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Arch Gen Psychiatry*, 60(6):565–71, Jun 2003.
- [5] Laura C Alonso and Robert L Rosenfield. Oestrogens and puberty. *Best Pract Res Clin Endocrinol Metab*, 16(1):13–30, Mar 2002.

- [6] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *Science*, 322(5903):881–8, Nov 2008.
- [7] Göran Arnqvist and Locke Rowe. *Sexual conflict*. Princeton University Press, Princeton, N.J., 2005.
- [8] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–9, May 2000.
- [9] Jeffrey A Bailey and Evan E Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–64, Jul 2006.
- [10] Jeffrey A Bailey, Zhiping Gu, Royden A Clark, Knut Reinert, Rhea V Samonte, Stuart Schwartz, Mark D Adams, Eugene W Myers, Peter W Li, and Evan E Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–7, Aug 2002.
- [11] Elizabeth Barrett-Connor. Commentary: Masculinity, femininity and heart disease. *Int J Epidemiol*, 36(3):621–2, Jun 2007.
- [12] U Becker and C Gluud. Sex, sex hormones and chronic liver diseases. *Dig Dis*, 9(1):9–16, 1991.

- [13] Jeffrey M Bhasin, Enakshi Chakrabarti, Dao-Quan Peng, Aneesh Kulkarni, Xi Chen, and Jonathan D Smith. Sex specific gene regulation and expression QTLs in mouse macrophages from a strain intercross. *PLoS One*, 3(1):e1435, 2008.
- [14] M. Björklund and J. C. Senar. Sex differences in autoimmune disease selection in the serin, *Serinus serinus*. *Journal of Evolutionary Biology*, 14(5):841–849, 2001.
- [15] E Blaak. Gender differences in fat metabolism. *Curr Opin Clin Nutr Metab Care*, 4(6):499–502, Nov 2001.
- [16] Ran Blekhman, Orna Man, Leslie Herrmann, Adam R Boyko, Amit Indap, Carolin Kosiol, Carlos D Bustamante, Kosuke M Teshima, and Molly Przeworski. Natural selection on genes that underlie human disease susceptibility. *Curr Biol*, 18(12):883–9, Jun 2008.
- [17] W.J. Bond and K.E. Maze. Survival costs and reproductive benefits of floral display in a sexually dimorphic dioecious shrub, *Leucadendron xanthoconus*. *Evolutionary Ecology*, 13(1):1–18, 1999.
- [18] Russell Bonduriansky and Stephen F Chenoweth. Intralocus sexual conflict. *Trends Ecol Evol*, 24(5):280–8, May 2009.
- [19] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore de Scienze Economiche e Commerciali de Firenze*, 8:3–62, 1936.

- [20] S J Boulton. Cellular functions of the BRCA tumour-suppressor proteins. *Biochem Soc Trans*, 34(Pt 5):633–45, Nov 2006.
- [21] Steven R Brant and Geoffrey C Nguyen. Is there a gender difference in the prevalence of Crohn’s disease or ulcerative colitis? *Inflamm Bowel Dis*, 14 Suppl 2:S2–3, Oct 2008.
- [22] Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J Natl Cancer Inst*, 98(19):1382–96, Oct 2006.
- [23] Elizabeth Buescher, Tilman Achberger, Idris Amusan, Anthony Gianini, Cherie Ochsenfeld, Ana Rus, Brett Lahner, Owen Hoekenga, Elena Yakubova, Jeffrey F Harper, Mary Lou Guerinot, Min Zhang, David E Salt, and Ivan R Baxter. Natural genetic variation in selected populations of *Arabidopsis thaliana* is associated with ionomic differences. *PLoS One*, 5(6):e11081, 2010.
- [24] P S Burgoyne. Genetic homology and crossing over in the X and Y chromosomes of mammals. *Hum Genet*, 61(2):85–90, 1982.
- [25] Merlin G Butler. Genomic imprinting disorders in humans: a mini-review. *J Assist Reprod Genet*, 26(9-10):477–86, 2009.
- [26] Andrea Camperio-Ciani, Francesca Corna, and Claudio Capiluppi. Evidence for maternally inherited factors favouring male homosexuality and promoting female fecundity. *Proc Biol Sci*, 271(1554):2217–21, Nov 2004.

- [27] Deborah Charlesworth and Brian Charlesworth. Sex chromosomes: evolution of the weird and wonderful. *Curr Biol*, 15(4):R129–31, Feb 2005.
- [28] J Chen, D P Silver, D Walpita, S B Cantor, A F Gazdar, G Tomlinson, F J Couch, B L Weber, T Ashley, D M Livingston, and R Scully. Stable interaction between the products of the BRCA1 and BRCA2 tumor suppressor genes in mitotic and meiotic cells. *Mol Cell*, 2(3):317–28, Sep 1998.
- [29] Joseph Cheung, Xavier Estivill, Razi Khaja, Jeffrey R MacDonald, Ken Lau, Lap-Chee Tsui, and Stephen W Scherer. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*, 4(4):R25, 2003.
- [30] A K Chippindale, J R Gibson, and W R Rice. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in *Drosophila*. *Proc Natl Acad Sci U S A*, 98(4):1671–5, Feb 2001.
- [31] Brian G Choi and Mary Ann McLaughlin. Why men’s hearts break: cardiovascular effects of sex steroids. *Endocrinol Metab Clin North Am*, 36(2):365–77, Jun 2007.
- [32] G A Churchill and R W Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–71, Nov 1994.
- [33] David Clayton and Hin-Tak Leung. An R package for analysis of whole-genome association studies. *Human heredity*, 64(1):45–51, 2007.

- [34] David G Clayton, Neil M Walker, Deborah J Smyth, Rebecca Pask, Jason D Cooper, Lisa M Maier, Luc J Smink, Alex C Lam, Nigel R Ovington, Helen E Stevens, Sarah Nutland, Joanna M M Howson, Malek Faham, Martin Moorhead, Hywel B Jones, Matthew Falkowski, Paul Hardenbol, Thomas D Willis, and John A Todd. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37(11):1243–6, Nov 2005.
- [35] Gavin C Conant and Andreas Wagner. Asymmetric sequence divergence of duplicate genes. *Genome Res*, 13(9):2052–8, Sep 2003.
- [36] Tim Connallon and L Lacey Knowles. Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends Genet*, 21(9):495–9, Sep 2005.
- [37] Karen N Conneely and Michael Boehnke. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet*, 81(6), Oct 2007.
- [38] David G Cox, Susan E Hankinson, and David J Hunter. No association between BRCA2 N372H and breast cancer risk. *Cancer Epidemiol Biomarkers Prev*, 14(5):1353–4, May 2005.
- [39] Robert M Cox and Ryan Calsbeek. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *Am Nat*, 173(2):176–87, Feb 2009.

- [40] Robert M Cox and Ryan Calsbeek. Cryptic sex-ratio bias provides indirect genetic benefits despite sexual conflict. *Science*, 328(5974):92–4, Apr 2010.
- [41] Troy Day and Russell Bonduriansky. Intralocus sexual conflict can drive the evolution of genomic imprinting. *Genetics*, 167(4):1537–46, Aug 2004.
- [42] Lynda F Delph, Janet L Gehring, Frank M Frey, A Michele Arntz, and Maureen Levri. Genetic constraints on floral evolution in a sexually dimorphic plant revealed by artificial selection. *Evolution*, 58(9):1936–46, Sep 2004.
- [43] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999.
- [44] Chunming Ding and Shengnan Jin. High-throughput methods for SNP genotyping. *Methods Mol Biol*, 578:245–54, 2009.
- [45] Paul F Doherty, Jr, Gabriele Sorci, J Andrew Royle, James E Hines, James D Nichols, and Thierry Boulinier. Sexual selection affects local extinction and turnover in bird communities. *Proc Natl Acad Sci U S A*, 100(10):5858–62, May 2003.
- [46] Greg L Drevenstedt, Eileen M Crimmins, Sarinnapha Vasunilashorn, and Caleb E Finch. The rise and fall of excess male infant mortality. *Proc Natl Acad Sci U S A*, 105(13):5016–21, Apr 2008.

- [47] Jo-Ann Eastwood and Lynn V Doering. Gender differences in coronary artery disease. *J Cardiovasc Nurs*, 20(5):340–51; quiz 352–3, 2005.
- [48] Richard F Edlich, Catherine L Cross, Courtney A Wack, Margot E Chase, Kant Y Lin, and William B Long, 3rd. Breast cancer and ovarian cancer genetics: an update. *J Environ Pathol Toxicol Oncol*, 27(4):245–56, 2008.
- [49] Hans Ellegren and John Parsch. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*, 8(9):689–698, 2007.
- [50] Xavier Estivill, Joseph Cheung, Miguel Angel Pujana, Kazuhiko Nakabayashi, Stephen W Scherer, and Lap-Chee Tsui. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (snps) correlate with segmental duplications in the human genome. *Hum Mol Genet*, 11(17):1987–95, Aug 2002.
- [51] James D Fackenthal and Olufunmilayo I Olopade. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat Rev Cancer*, 7(12):937–48, Dec 2007.
- [52] Peter A Fasching, Simon Gayther, Leigh Pearce, Joellen M Schildkraut, Ellen Goode, Falk Thiel, Georgia Chenevix-Trench, Jenny Chang-Claude, Shan Wang-Gohrke, Susan Ramus, Paul Pharoah, Andrew Berchuck, and OCAC (Ovarian Cancer Association Consortium). Role of genetic polymorphisms and ovarian cancer susceptibility. *Mol Oncol*, 3(2):171–81, Apr 2009.

- [53] Andrew P Feinberg, Hengmi Cui, and Rolf Ohlsson. DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. *Semin Cancer Biol*, 12(5):389–98, Oct 2002.
- [54] L M Field, F Lyko, M Mandrioli, and G Prantera. DNA methylation in insects. *Insect Mol Biol*, 13(2):109–15, Apr 2004.
- [55] Anthony C Fiumera, Bethany L Dumont, and Andrew G Clark. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. *Genetics*, 176(2):1245–60, Jun 2007.
- [56] Antonia Flaquer, Gudrun A Rappold, Thomas F Wienker, and Christine Fischer. The human pseudoautosomal regions: a review for genetic epidemiologists. *Eur J Hum Genet*, 16(7):771–9, Jul 2008.
- [57] Katharina Foerster, Tim Coulson, Ben C Sheldon, Josephine M Pemberton, Tim H Clutton-Brock, and Loeske E B Kruuk. Sexually antagonistic genetic variation for fitness in red deer. *Nature*, 447(7148):1107–10, Jun 2007.
- [58] Brad Foley, Stephen F Chenoweth, Sergey V Nuzhdin, and Mark W Blows. Natural genetic variation in cuticular hydrocarbon expression in male and female *Drosophila melanogaster*. *Genetics*, 175(3):1465–77, Mar 2007.
- [59] Matthew L Freedman, Kathryn L Penney, Daniel O Stram, Loïc Le Marchand, Joel N Hirschhorn, Laurence N Kolonel, David Altshuler, Brian E

- Henderson, and Christopher A Haiman. Common variation in BRCA2 and breast cancer risk: a haplotype-based analysis in the Multiethnic Cohort. *Hum Mol Genet*, 13(20):2431–41, Oct 2004.
- [60] D Freije, C Helms, M S Watson, and H Donis-Keller. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science*, 258(5089):1784–7, Dec 1992.
- [61] J Gadau, R E Page, and J H Werren. The genetic basis of the interspecific differences in wing size in *Nasonia* (Hymenoptera; Pteromalidae): major quantitative trait loci and epistasis. *Genetics*, 161(2):673–84, Jun 2002.
- [62] E A Gale and K M Gillespie. Diabetes and gender. *Diabetologia*, 44(1):3–15, Jan 2001.
- [63] Xiaoyi Gao, Lewis C Becker, Diane M Becker, Joshua D Starmer, and Michael A Province. Avoiding the high bonferroni penalty in genome-wide association studies. *Genet Epidemiol*, 34(1):100–5, Jan 2010.
- [64] Xiaoyi Gao, Joshua Starmer, and Eden R Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*, 32(4):361–9, May 2008.
- [65] Montserrat García-Closas, Kathleen M Egan, Polly A Newcomb, Louise A Brinton, Linda Titus-Ernstoff, Stephen Chanock, Robert Welch, Jolanta Lissowska, Beata Peplonska, Neonila Szeszenia-Dabrowska, Witold Zatonski, Alicja Bardin-Mikolajczak, and Jeffery P Struwing. Polymorphisms

- in DNA double-strand break repair genes and risk of breast cancer: two population-based studies in USA and Poland, and meta-analyses. *Hum Genet*, 119(4):376–88, May 2006.
- [66] Greg Gibson. Hints of hidden heritability in gwas. *Nat Genet*, 42(7):558–60, Jul 2010.
- [67] Florian Gnad and John Parsch. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. *Bioinformatics*, 22(20):2577–9, Oct 2006.
- [68] Kevin L Gunderson, Frank J Steemers, Grace Lee, Leo G Mendoza, and Mark S Chee. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, May 2005.
- [69] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456, Apr 2009.
- [70] Daniel L. Hartl. *A primer of population genetics*, page 74. Sinauer Associates, Inc., 3rd edition, 2000.
- [71] C S Healey, A M Dunning, M D Teare, D Chase, L Parker, J Burn, J Chang-Claude, A Mannermaa, V Kataja, D G Huntsman, P D Pharoah, R N Luben, D F Easton, and B A Ponder. A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. *Nat Genet*, 26(3):362–4, Nov 2000.

- [72] Louise Hosking, Sheena Lumsden, Karen Lewis, Astrid Yeo, Linda McCarthy, Aruna Bansal, John Riley, Ian Purvis, and Chun-Fang Xu. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet*, 12(5):395–9, May 2004.
- [73] G A Huttley and S R Wilson. Testing for concordant equilibria between population samples. *Genetics*, 156(4):2127–35, Dec 2000.
- [74] Francesca Iemmola and Andrea Camperio Ciani. New evidence of genetic factors influencing sexual orientation in men: female fecundity increase in the maternal line. *Arch Sex Behav*, 38(3):393–9, Jun 2009.
- [75] Online Mendelian Inheritance in Man.
<http://www.ncbi.nlm.nih.gov/omim>.
- [76] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, Oct 2005.
- [77] Peter Kang, Shivaani Mariapun, Sze Yee Phuah, Linda Shushan Lim, Jianjun Liu, Sook-Yee Yoon, Meow Keong Thong, Nur Aishah Mohd Taib, Cheng Har Yip, and Soo-Hwang Teo. Large BRCA1 and BRCA2 genomic rearrangements in Malaysian high risk breast-ovarian cancer families. *Breast Cancer Res Treat*, 124(2):579–84, Nov 2010.
- [78] Brian A Kennedy, Wenqing Gao, Tim H-M Huang, and Victor X Jin. HRT-BLDb: an informative data resource for hormone receptors target binding loci. *Nucleic Acids Res*, 38(Database issue):D676–81, Jan 2010.

- [79] W Jason Kennington, Ary A Hoffmann, and Linda Partridge. Mapping regions within cosmopolitan inversion *In(3R)Payne* associated with natural variation in body size in *drosophila melanogaster*. *Genetics*, 177(1):549–56, Sep 2007.
- [80] W James Kent. BLAT—the BLAST-like alignment tool. *Genome Res*, 12(4):656–64, Apr 2002.
- [81] S Khosla, L J Melton, 3rd, E J Atkinson, W M O’Fallon, G G Klee, and B L Riggs. Relationship of serum sex steroid levels and bone turnover markers with bone mineral density in men and women: a key role for bioavailable estrogen. *J Clin Endocrinol Metab*, 83(7):2266–74, Jul 1998.
- [82] J F Kidwell, M T Clegg, F M Stewart, and T Prout. Regions of stable equilibria for models of differential selection in the two sexes under random mating. *Genetics*, 85(1):171–83, Jan 1977.
- [83] Mary-Claire King, Joan H Marks, Jessica B Mandell, and New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, 302(5645):643–6, Oct 2003.
- [84] C Köhler and I Weinhofer-Molisch. Mechanisms and evolution of genomic imprinting in plants. *Heredity*, 105(1):57–63, Jul 2010.
- [85] The Jackson Laboratory. <http://www.jax.org/>.

- [86] Kristin R Lamont and Donald J Tindall. Androgen regulation of gene expression. *Adv Cancer Res*, 107:137–62, 2010.
- [87] H Lan, C M Kendzierski, J D Haag, L A Shepel, M A Newton, and M N Gould. Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. *Genetics*, 157(1):331–9, Jan 2001.
- [88] R. Lande. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution*, 34(2):292–305, 1980.
- [89] R Lande and M Kirkpatrick. Ecological speciation by sexual selection. *J Theor Biol*, 133(1):85–98, Jul 1988.
- [90] Cathy C Laurie, Kimberly F Doheny, Daniel B Mirel, Elizabeth W Pugh, Laura J Bierut, Tushar Bhangale, Frederick Boehm, Neil E Caporaso, Marilyn C Cornelis, Howard J Edenberg, Stacy B Gabriel, Emily L Harris, Frank B Hu, Kevin B Jacobs, Peter Kraft, Maria Teresa Landi, Thomas Lumley, Teri A Manolio, Caitlin McHugh, Ian Painter, Justin Paschall, John P Rice, Kenneth M Rice, Xiuwen Zheng, Bruce S Weir, and GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*, 34(6):591–602, Sep 2010.
- [91] Richard J Leach, Susan C Craigmile, Sara A Knott, John L Williams, and Elizabeth J Glass. Quantitative trait loci for variation in immune response to a Foot-and-Mouth Disease virus peptide. *BMC Genet*, 11:107, 2010.

- [92] Suzanne M Leal. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol*, 29(3):204–14, Nov 2005.
- [93] The Scopus literature database. <http://www.scopus.com/>.
- [94] A D Long and C H Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res*, 9(8):720–31, Aug 1999.
- [95] Dimitrios Loukovitis, Elena Sarropoulou, Costas S Tsigenopoulos, Costas Batargias, Antonios Magoulas, Apostolos P Apostolidis, Dimitrios Chatzipilis, and Georgios Kotoulas. Quantitative trait loci involved in sex determination and body growth in the gilthead sea bream (*Sparus aurata* L.) through targeted genome scan. *PLoS One*, 6(1):e16599, 2011.
- [96] Faidon Magkos and Bettina Mittendorfer. Gender differences in lipid metabolism and the effect of obesity. *Obstet Gynecol Clin North Am*, 36(2):245–65, vii, Jun 2009.
- [97] Judith E Mank. Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *Am Nat*, 173(2):141–50, Feb 2009.
- [98] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, Jun 2010.

- [99] Hajime Matsuzaki, Halina Loi, Shoulian Dong, Ya-Yu Tsai, Joy Fang, Jane Law, Xiaojun Di, Wei-Min Liu, Geoffrey Yang, Guoying Liu, Jing Huang, Giulia C Kennedy, Thomas B Ryder, Gregory A Marcus, P Sean Walsh, Mark D Shriver, Jennifer M Puck, Keith W Jones, and Rui Mei. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res*, 14(3):414–25, Mar 2004.
- [100] Steven A McCarroll and David M Altshuler. Copy-number variation and association studies of human disease. *Nat Genet*, 06 2007.
- [101] Steven A McCarroll, Finny G Kuruvilla, Joshua M Korn, Simon Cawley, James Nemesh, Alec Wysoker, Michael H Shapero, Paul I W de Bakker, Julian B Maller, Andrew Kirby, Amanda L Elliott, Melissa Parkin, Earl Hubbell, Teresa Webster, Rui Mei, James Veitch, Patrick J Collins, Robert Handsaker, Steve Lincoln, Marcia Nizzari, John Blume, Keith W Jones, Rich Rava, Mark J Daly, Stacey B Gabriel, and David Altshuler. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40(10):1166–74, Oct 2008.
- [102] P A McCombe, J M Greer, and I R Mackay. Sexual dimorphism in autoimmune disease. *Curr Mol Med*, 9(9):1058–79, Dec 2009.
- [103] Stuart F McDaniel. Genetic correlations do not constrain the evolution of sexual dimorphism in the moss *Ceratodon purpureus*. *Evolution*, 59(11):2353–61, Nov 2005.

- [104] Colin D Meiklejohn, John Parsch, José M Ranz, and Daniel L Hartl. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc Natl Acad Sci U S A*, 100(17):9894–9, Aug 2003.
- [105] J. Merilä, B. C. Sheldon, and H. Ellegren. Antagonistic natural selection revealed by molecular sex identification of nestling collared flycatchers. *Molecular ecology*, 6(12):1167–1175, 1997. Cited By (since 1996): 41.
- [106] Kelly Metcalfe, Henry T Lynch, Parviz Ghadirian, Nadine Tung, Ivo Olivotto, Ellen Warner, Olufunmilayo I Olopade, Andrea Eisen, Barbara Weber, Jane McLennan, Ping Sun, William D Foulkes, and Steven A Narod. Contralateral breast cancer in BRCA1 and BRCA2 mutation carriers. *J Clin Oncol*, 22(12):2328–35, Jun 2004.
- [107] Jason H Moore, Folkert W Asselbergs, and Scott M Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–55, Feb 2010.
- [108] T Moore and D Haig. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet*, 7(2):45–9, Feb 1991.
- [109] E H Morrow, A D Stewart, and W R Rice. Assessing the extent of genome-wide intralocus sexual conflict via experimentally enforced gender-limited selection. *J Evol Biol*, 21(4):1046–54, Jul 2008.

- [110] Benjamin M Neale and Shaun Purcell. The positives, protocols, and perils of genome-wide association. *Am J Med Genet B Neuropsychiatr Genet*, 147B(7):1288–94, Oct 2008.
- [111] Derrick L Nehrenberg, Shiliang Wang, Ryan J Buus, James Perkins, Fernando Pardo-Manuel de Villena, and Daniel Pomp. Genomic mapping of social behavior traits in a F2 cross derived from mice selectively bred for high aggression. *BMC Genet*, 11:113, 2010.
- [112] Carole Ober, Dagan A Loisel, and Yoav Gilad. Sex-specific genetic architecture of human disease. *Nat Rev Genet*, 9(12):911–22, Dec 2008.
- [113] Jacqueline E Oliver and Alan J Silman. Why are women predisposed to autoimmune rheumatic diseases? *Arthritis Res Ther*, 11(5):252, 2009.
- [114] Nancy J Olsen and William J Kovacs. Hormones, pregnancy, and rheumatoid arthritis. *J Gend Specif Med*, 5(4):28–37, 2002.
- [115] Roman Pahl and Helmut Schäfer. PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics*, 26(17):2093–100, Sep 2010.
- [116] Demosthenes B Panagiotakos. Value of p -value in biomedical research. *Open Cardiovasc Med J*, 2:97–9, 2008.
- [117] Michael Parisi, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. Paucity of

- genes on the *Drosophila* X chromosome showing male-biased expression. *Science*, 299(5607):697–700, Jan 2003.
- [118] G A Parker and L Partridge. Sexual conflict and speciation. *Philos Trans R Soc Lond B Biol Sci*, 353(1366):261–74, Feb 1998.
- [119] GA Parker. Sexual selection and sexual conflict. *Sexual selection and reproductive competition in insects*. Academic Press, New York, pages 123–166, 1979.
- [120] Manus M Patten and David Haig. Reciprocally imprinted genes and the response to selection on one sex. *Genetics*, 179(3):1389–94, Jul 2008.
- [121] Manus M Patten and David Haig. Maintenance or loss of genetic variation under sexual and parental antagonism at a sex-linked locus. *Evolution*, 63(11):2888–95, Nov 2009.
- [122] Elizabeth Pennisi. The biology of genomes. disease risk links to gene regulation. *Science*, 332(6033):1031, May 2011.
- [123] J. Perkel. SNP genotyping: six technologies that keyed a revolution. *Nature Methods*, 5(5):447–454, 2008.
- [124] Alison Pischedda and Adam K Chippindale. Intralocus sexual conflict diminishes the benefits of sexual selection. *PLoS Biol*, 4(11):e356, Nov 2006.

- [125] François Pompanon, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet*, 6(11):847–59, Nov 2005.
- [126] Michael L Power and Jay Schulkin. Sex differences in fat storage, fat metabolism, and the health risks from obesity: possible evolutionary origins. *Br J Nutr*, 99(5):931–40, May 2008.
- [127] Bedhomme S.-Day T. Chippindale A.K. Prasad, N.G. An evolutionary cost of separate genders revealed by male-limited evolution. *American Naturalist*, 169(1):29–37, 2007. cited By (since 1996) 33.
- [128] D. K. Price and N. T. Burley. Constraints on the evolution of attractive traits: Selection in male and female zebra finches. *American Naturalist*, 144(6):908–934, 1994. Cited By (since 1996): 47.
- [129] The Ensembl project. <http://www.ensembl.org>.
- [130] S Purcell, S S Cherny, and P C Sham. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–50, Jan 2003.
- [131] Li-Xin Qiu, Lei Yao, Kai Xue, Jian Zhang, Chen Mao, Bo Chen, Ping Zhan, Hui Yuan, and Xi-Chun Hu. BRCA2 N372H polymorphism and breast cancer susceptibility: a meta-analysis involving 44,903 subjects. *Breast Cancer Res Treat*, Feb 2010.

- [132] José M Ranz, Cristian I Castillo-Davis, Colin D Meiklejohn, and Daniel L Hartl. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science*, 300(5626):1742–5, Jun 2003.
- [133] T Rhen. Sex-limited mutations and the evolution of sexual dimorphism. *Evolution*, 54(1):37–43, Feb 2000.
- [134] Treva K Rice, Nicholas J Schork, and D C Rao. Methods for handling multiple testing. *Adv Genet*, 60:293–308, 2008.
- [135] William R Rice and Adam K Chippindale. The evolution of hybrid infertility: perpetual coevolution between gender-specific and sexually antagonistic genes. *Genetica*, 116(2-3):179–88, Nov 2002.
- [136] W.R. Rice. Sex chromosomes and the evolution of sexual dimorphism. *Evolution*, 38(4):735–742, 1984. cited By (since 1996) 293.
- [137] W.R. Rice. The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. *Evolution*, 41(4):911–914, 1987.
- [138] W.R. Rice. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature*, 381(6579):232–234, 1996. cited By (since 1996) 497.
- [139] M D Ritchie, L W Hahn, N Roodi, L R Bailey, W D Dupont, F F Parl, and J H Moore. Multifactor-dimensionality reduction reveals high-order

- interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69(1):138–47, Jul 2001.
- [140] Locke Rowe and Troy Day. Detecting sexual conflict and sexually antagonistic coevolution. *Philos Trans R Soc Lond B Biol Sci*, 361(1466):277–85, Feb 2006.
- [141] R Sachidanandam, D Weissman, S C Schmidt, J M Kakol, L D Stein, G Marth, S Sherry, J C Mullikin, B J Mortimore, D L Willey, S E Hunt, C G Cole, P C Coggill, C M Rice, Z Ning, J Rogers, D R Bentley, P Y Kwok, E R Mardis, R T Yeh, B Schultz, L Cook, R Davenport, M Dante, L Fulton, L Hillier, R H Waterston, J D McPherson, B Gilman, S Schaffner, W J Van Etten, D Reich, J Higgins, M J Daly, B Blumenstiel, J Baldwin, N Stange-Thomann, M C Zody, L Linton, E S Lander, D Altshuler, and International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33, Feb 2001.
- [142] Georgia Salanti, Georgia Amountza, Evangelia E Ntzani, and John P A Ioannidis. Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *Eur J Hum Genet*, 13(7):840–8, Jul 2005.
- [143] M.J. Schervish. p values: What they are and what they are not. *The American Statistician*, 50(3), 1996.

- [144] S K Sharan, M Morimatsu, U Albrecht, D S Lim, E Regel, C Dinh, A Sands, G Eichele, P Hasty, and A Bradley. Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature*, 386(6627):804–10, Apr 1997.
- [145] Lisa M Shulman. Gender differences in Parkinson’s disease. *Gend Med*, 4(1):8–18, Mar 2007.
- [146] J Gustav Smith and Christopher Newton-Cheh. Genome-wide association study in humans. *Methods Mol Biol*, 573:231–58, 2009.
- [147] Amanda B Spurdle, John L Hopper, Xiaoqing Chen, Gillian S Dite, Jisheng Cui, Margaret R E McCredie, Graham G Giles, Sarah Ellis-Steinborner, Deon J Venter, Beth Newman, Melissa C Southey, and Georgia Chenevix-Trench. The BRCA2 372 HH genotype is associated with risk of breast cancer in Australian women under age 60 years. *Cancer Epidemiol Biomarkers Prev*, 11(4):413–6, Apr 2002.
- [148] Stephen C Stearns, Sean G Byars, Diddahally R Govindaraju, and Douglas Ewbank. Measuring selection in contemporary human populations. *Nat Rev Genet*, 11(9):611–22, Sep 2010.
- [149] Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen, and Kevin L Gunderson. Whole-genome genotyping with the single-base extension assay. *Nat Methods*, 3(1):31–3, Jan 2006.

- [150] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64(3):479–498, 2002.
- [151] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, 2003.
- [152] J P Struewing, P Hartge, S Wacholder, S M Baker, M Berlin, M McAdams, M M Timmerman, L C Brody, and M A Tucker. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med*, 336(20):1401–8, May 1997.
- [153] Erik I Svensson, Andrew G McAdam, and Barry Sinervo. Intralocus sexual conflict over immune defense, gender load, and sex-specific signaling in a natural lizard population. *Evolution*, 63(12):3124–35, Dec 2009.
- [154] M D Teare, A Cox, J Shorto, C Anderson, D T Bishop, and C Cannings. Heterozygote excess is repeatedly observed in females at the BRCA2 locus N372H. *J Med Genet*, 41(7):523–8, Jul 2004.
- [155] Yik Y Teo. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol*, 19(2):133–143, 2008.
- [156] Ammarin Thakkinstian, Patrick McElduff, Catherine D’Este, David Duffy, and John Attia. A method for meta-analysis of molecular association studies. *Stat Med*, 24(9):1291–306, May 2005.

- [157] Paul D Thomas, Michael J Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. PANTHER: a library of protein families and sub-families indexed by function. *Genome Res*, 13(9):2129–41, Sep 2003.
- [158] Paul D Thomas, Anish Kejariwal, Michael J Campbell, Huaiyu Mi, Karen Diemer, Nan Guo, Istvan Ladunga, Betty Ulitsky-Lazareva, Anushya Muruganujan, Steven Rabkin, Jody A Vandergriff, and Olivier Doremieux. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*, 31(1):334–41, Jan 2003.
- [159] P Tripputi, D Cigognini, S Bianchi, and L Fedele. Different allelic distribution of a single SNP between sexes in humans. *Biochem Genet*, 46(11-12):733–6, Dec 2008.
- [160] Circe Tsui, Laura E Coleman, Jacquelyn L Griffith, E Andrew Bennett, Summer G Goodson, Jason D Scott, W Stephen Pittard, and Scott E Devine. Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res*, 31(16):4910–6, Aug 2003.
- [161] G Sander van Doorn. Intralocus sexual conflict. *Ann N Y Acad Sci*, 1168:52–71, Jun 2009.
- [162] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides,

R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng,
 L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Ga-
 bor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick,
 N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller,
 R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern,
 S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington,
 J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill,
 I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco,
 P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong,
 Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum,
 Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Mil-
 shina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B
 Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang,
 J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang,
 H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert,
 S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An,
 A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam,
 A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R De-
 silets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann,
 B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck,
 T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love,
 F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy,

B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Rendon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lipert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Foster, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 2001.

- [163] Beatriz Vicoso and Brian Charlesworth. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet*, 7(8):645–53, Aug 2006.
- [164] Jose Viña and Ana Lloret. Why women have more Alzheimer’s disease than men: gender and mitochondrial toxicity of amyloid-beta peptide. *J*

Alzheimers Dis, 20 Suppl 2:S527–33, 2010.

- [165] Rhonda Voskuhl. Sex differences in autoimmune diseases. *Biol Sex Differ*, 2(1):1, 2011.
- [166] Ping Wang, Richard F Lyman, Svetlana A Shabalina, Trudy F C Mackay, and Robert R H Anholt. Association of polymorphisms in odorant-binding protein genes with variation in olfactory response to benzaldehyde in *Drosophila*. *Genetics*, 177(3):1655–65, Nov 2007.
- [167] Richard M Watanabe. Statistical issues in gene association studies. *Methods Mol Biol*, 700:17–36, 2011.
- [168] Michael E Weale. Quality control for genome-wide association studies. *Methods Mol Biol*, 628:341–72, 2010.
- [169] The Wellcome Trust Case-Control Consortium website.
<http://www.wtccc.org.uk/>.
- [170] Consortium Wellcome Trust Case Control. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.
- [171] Thomas M Williams and Sean B Carroll. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nat Rev Genet*, 10(11):797–804, Nov 2009.

- [172] I W Wilson, C L Schiff, D E Hughes, and S C Somerville. Quantitative trait loci analysis of powdery mildew disease resistance in the *Arabidopsis thaliana* accession Kashmir-1. *Genetics*, 158(3):1301–9, Jul 2001.
- [173] S J Winters, E Talbott, D S Guzick, J Zborowski, and K P McHugh. Serum testosterone levels decrease in middle age in women with the polycystic ovary syndrome. *Fertil Steril*, 73(4):724–9, Apr 2000.
- [174] A M Wolfe, J H Kellgren, and A T Masi. The epidemiology of rheumatoid arthritis: a review. II. Incidence and diagnostic criteria. *Bull Rheum Dis*, 19(3):524–9, Nov 1968.
- [175] Andrew J Wood and Rebecca J Oakey. Genomic imprinting in mammals: emerging themes and established theories. *PLoS Genet*, 2(11):e147, Nov 2006.
- [176] R Wooster, G Bignell, J Lancaster, S Swift, S Seal, J Mangion, N Collins, S Gregory, C Gumbs, and G Micklem. Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559):789–92, 1995.
- [177] G F Wooten, L J Currie, V E Bovbjerg, J K Lee, and J Patrie. Are men at greater risk for Parkinson’s disease than women? *J Neurol Neurosurg Psychiatry*, 75(4):637–9, Apr 2004.
- [178] Jianfeng Xu, Aubrey Turner, Joy Little, Eugene R Bleecker, and Deborah A Meyers. Positive results in association studies are associated with departure

from Hardy-Weinberg equilibrium: hint for genotyping error? *Hum Genet*, 111(6):573–4, Dec 2002.

- [179] Andreas Ziegler. Genome-wide association studies: quality control and population-based measures. *Genet Epidemiol*, 33 Suppl 1:S45–50, 2009.