THE FAMILY STUDY OF HIGH MYOPIA:

ASSOCIATION STUDIES

Tetyana Zayats
School of Optometry and Vision Sciences

Submitted to Cardiff University for the Degree of Doctor of Philosophy

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGEMENTS

I would like to sincerely thank my supervisor Dr. Jeremy A. Guggenheim. His broad knowledge, remarkable logic and concepts have been of great value to me. He has guided me through the years of my PhD with exceptional patience and understanding. I am truly honored to know him not only as an extraordinary teacher, but also as a good friend.

It is my pleasure to thank my advisor, Dr. Julie Albon, for her greatly appreciated suggestions. Her detailed comments on my reports have helped me to discover and explore interesting angles of my project.

I owe gratitude to Dr. Rosalind Creer for being a wonderful team-mate and friend. She introduced me not only to the beauty of the labwork, assembling and mailing our information packs, placing and collecting orders; but also to the ever important social life of PhD students in Cardiff.

I am grateful to Dr. Jonathan Ericshen for his esteemed support. His remarks and questions I received during meetings of the Visual Neuroscience and Molecular Biology group have always been thought-provoking and encouraging for me.

During this work I have collaborated with the genetic group of the Duke University, USA. I would like to extend my profound thanks to Prof. Terri L. Young, Dr.Yi-Ju Li, Tammy Yanovitch and Ravikanth Metlapally for their most highly respected help and advices.

Special acknowledgements go to the participants of the Family Study of Myopia, without whom this project would not have been possible.

I am genuinely thankful to Dr. Hywel Williams for his proficient help with microstellite genotyping and to Dr. Beate Glaser for her expertise on analysis of imprinting.

I would like to express warm thanks to Susan Hobbs, Leanne Jones and Stephen Morgan for their kind help in secretarial work and financial issues; and to Andy, Gregg, Rob and Jon for always being ready to help in any way they could. Many thanks are due to my colleagues from the School of Optometry and Vision Sciences for creating a friendly and inspiring atmosphere to work in. The School has also financed my project, which I deeply acknowledge.

I am indebted to many of my friends, especially to Christian, Ankush, Paul, Clark, Sally, Rob, Melody, Donna, Magda, Miguel, Yaiza, Franziska and Li. They have always been there for me throughout years and I feel extremely fortunate to have their support.

Words fail me to express my affectionate gratitude to my family. Their never-ending love, understanding and protection have laid the foundation of my character. Their example of hard work and their passion for knowledge have always been and will always be an inspiration for my scientific ambitions.

SUMMARY

High myopia (more severe than –6.00 Diopters) is one of the leading causes of blindness and vision impairment in the world. Its prevalence has rapidly been growing and the estimated number of myopic people worldwide is expected to be ~2.5 billion by the year 2020.

My experimental work covered three topics: (1) characterization of the quantity and quality of mouthwash-extracted DNA; (2) genetic association studies, and (3) evaluation of an imprinting effect in high myopia.

Mouthwash-derived DNA is an important source of human DNA for large-scale genetic studies. Thus, potential methods of DNA quantification (spectrophotometry, fluorometry, gel electrophoresis and qPCR) and quality assessment (gel electrophoresis and PCR) were evaluated. Regarding DNA quantification methods, fluorometry compared favorably to the gold-standard qPCR. DNA quality assessments revealed that ~10% of collected buccal DNA samples were severely degraded – a phenomenon that was shown to be partly subject-specific.

Myopia association studies were performed for: genes in MYP regions, the myocilin gene, the collagen type I alpha 1 gene and the collagen type II alpha 1 gene. These genes have been linked to myopia because of their function and/or previous positive findings. All tests were performed on a combined dataset of complex high myopia pedigrees and cases/controls, applying likelihood ratio statistics and Bonferroni correction to account for multiple testing. The results suggested that none of the genes examined have an important influence on susceptibility to high myopia.

There is greater resemblance of refractive error between siblings than between parents and offspring, implying the possibility of imprinting in the aetiology of myopia. Thus, tests for imprinting were performed on "trio" pedigrees, applying Z-score and T^2-test statistics and permutation to account for multiple testing. The results tentatively suggested that parent-of-origin effects and/or by maternal effects contribute to myopia development.

List of Figures

List of Tables

INTRODUCTION

Originating from Aristotle, reports on myopia (shortsightedness) are known from about the last 2000 years [1]. Yet, centuries later, it still cannot be fully interpreted. Shortsightedness has been the subject of much discussion and has inspired a number of theories. However, there is still no satisfactory explanation for this condition.

The prevalence of myopia has rapidly been growing in some parts of the world [2-4]. The estimated number of 1.6 billion myopic people worldwide is expected to increase to approximately 2.5 billion by the year 2020 [2]. The World Health Organization has put myopia among the leading causes of blindness and vision impairment in the world [3].

Aiming to help the understanding of shortsightedness, the project named "The Family Study of Myopia" was started about 10 years ago. In 2005, I joined the study as a PhD student and this thesis is the result of the work I have done under the guidance of my supervisor and in collaboration with 4 other myopia research groups. My project was focused on the genetic association studies and the scope of my activities included (1) subject recruitment (section 2.1); (2) processing (section 2.2) and assessing of participants' DNA (chapter III.); (3) microsatllite genotyping (section 2.3) and (4) the actual statistical tests of various nature - replication, genome-wide association and imprinting (chapters IV-VII).

This work concerns high myopia, that is more severe than -6.0 Doiptres (section 2.4.4), as the higher degrees of the condition represent a particular threat by inducing pathological changes that can lead to blindness (section 1.2.3).

The aim of the first chapter is to provide the reader with background information on myopia and genetics that, in my opinion, is essential for the understanding of the later chapters. The following section on Materials and Methods introduces all of the techniques that I have used and is intended to serve as a reference for the technical parts later on. Finally, chapters III to VII report my findings.

There are many papers published on the subject of myopia. While it is often difficult to compare different studies, I have decided to include scientifically weaker ones, as well as the stronger ones, to show the reader the points of contradiction in the literature on myopia and the variety of theories intended to explain the mechanism of shortsightedness.

My first task was subject recruitment and DNA sample collection in the form of mouthwash (sections 2.1), followed by DNA extraction (section 2.2.1). As The Family Study of Myopia was an established and running project when I joined, there already was a databade of the previously recruited subjects and their extracted DNA. My contribution was the collection of additional 19 families (150 subjects), 60 cases and 111 controls as well as their DNA.

The next section (section 2.2.2) describes the four methods - spectrophotometry, fluorometry, gel electrophoresis and polymerase chain reaction - that I utilized to perform a quantity and quality control on DNA extracted by me from the new samples that I have collected, as well as on the DNA that was already in the database of The Family Study of Myopia. This work is detailed in chapter III. Later on, I applied the developed quality control method to all mouthwash-extracted DNA that was genotyped for the association analyses.

This study concerns two types of DNA polymorphisms: microsatellites and SNPs. The first variants were genotyped by me using the technique described in section 2.3.1, while SNPs were genotyped by various companies (each of which is specified in an appropriate part of the thesis). Section 2.3.2 gives an overview of the SNP genotyping techniques. Once the variants were genotyped, I performed a quality check (section 2.3.3) on all data independently of whether it was obtained by me or by a genotyping company.

My work also involved the performance of a number of statistical analyses, that are described in section 2.4 of chapter II on materials and methods. This section first characterize such methods as generalized linear model and analyses of categorical outcomes, that I used to draw conclusions from experiments performed on the mouthwash-extracted DNA (chapter III.). Further sections on likelihood, statistics of

association and the issue of multiple testing intend to give the reader an idea about how the genetic softwares, I have utilized in chapters IV-VII, work and how did I decide whether each specific result was genuine or a false positive due to chance.

Finally, chapters IV-VII are designated to the association analyses I carried out to gain information about the genetic background of severe myopia. The description of the work is in the subsequent chapters on two replication studies (Myocilin and Collagen genes), one genome-wide association study and, finally, an examination of the possible epigenetic impact on high myopia. All these analyses were performed by me. The test of collagen gene (chapter VI.) was carried out on the subjects recruited within The Family Study of Myopia only; while the genome-wide association (chapter V.), the myocilin gene replication (chapter IV.) and test of epigenic effect (chapter VII.) involved participants collected by our collaborators. In addition, the relationship between collagen genes and high myopia was also examined by the research group of Prof. T. Young. I included their findings along with my own in chapter VI. to allow a comparison.

CHAPTER I.

MYOPIA AND GENETICS BACKGROUND

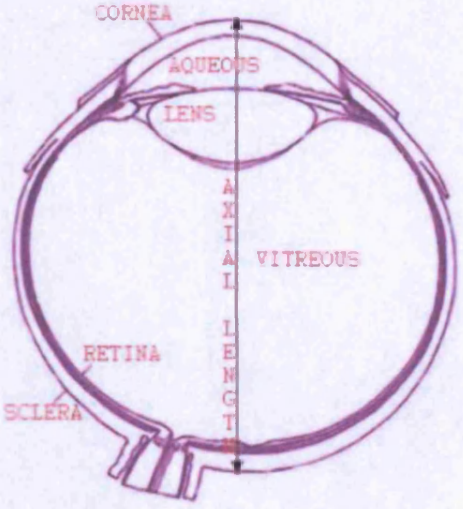
1.1 Myopia

1.1.1 Myopia as a refractive error

1.1.1.1 Ocular Components and Myopia

The human eye is a delicately balanced visual system comprising several components (Figure 1.1). The interaction of these elements defines an ocular refractive status, with reference to an eye in which parallel rays of light from infinity are brought to a focus, with relaxed accommodation [4]. In an emmetropic eye, this focus is projected exactly upon the retina, while in ametropia (refractive error) it is in front (myopia) or beyond (hyperopia) the retina.

Figure1.1 The Structure of the Human Eye
(Adopted from Mouroulis [5])



Refractive surfaces (refractive power):
- Cornea
- Lens

Refractive Indices:
- Aqueous (anterior chamber)
- Vitreous (vitreous chamber)

Linear Ditances:
- Anterior chamber depth
- Vitreous chamber depth
- Axial length

Emmetropia is the result of the development and maintenance of a precise optical arrangement and structure of the eye (emmetropization), any imbalance in which leads to refractive error. Ametropia arises if either refractive power (cornea, lens) or axial length deviates from the optimal (normal) state, while effects of refractive indices (aqueous, lenticular and vitreous) are usually invariant [6]. Thus, parallel rays of light can be focused in front of the retina (myopia) as a consequence of too great a refractive power or too great an axial length [7].

Ocular axial length (Figure 1.2) has a high correlation with ametropia [8] and seems to be the major factor in the development [9] and progression of myopia [9, 10]. The radius of curvature of the cornea (Figure 1.3) and the power of the crystalline lens (Figure 1.4), on the other hand, appear to show a much weaker correlation with ametropia [8, 10].

Nonetheless, refraction is correlated with the combined effects of ocular power and axial length [10], engaging cornea, lens and the length of the eye together, in a complex interplay.

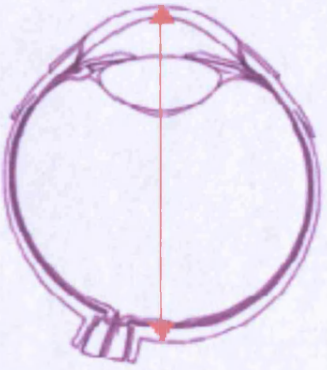


Figure 1.2 The Axial Length
(Adopted from Mouroulis [5])

Length variation with age [11]:
- at birth: 17mm
- in adulthood: 24mm

Length variation with ametropia [12]:
- emmetropia: 24mm
- myopia: >26mm

Figure 1.3 The Cornea
(Adopted from Mouroulis [5])

Refractive power [11]:
- at birth: 51 Dioptres
- in adulthood: 43 Dioptres

Figure 1.4 The Lens
(Adopted from Mouroulis [5])

Refractive power [11]:
- at birth: 34 Dioptres
- in adulthood: 18 Dioptres

1.1.1.2 Emmetropization and Myopia

Humans are usually born with hyperopic errors - the eye is too short for the optical power of cornea and lens. Thus, during development in the early years of life the eye elongates (emmetropization) to match the power of its optical components [13, 14]. Emmetropia is usually reached by about 6-7 years of age [15]. The failure to reach or maintain emmetropia results in ametropia.

Generally, there are two phases of emmetropization: a "rapid" phase of fast axial growth during infancy and a "slow" phase during school years [11, 16]. Most emmetropization takes place during the rapid growth phase, especially between 3 and 12 months of age. The changes of the eye in this phase are axial lengthening and the loss of the power of the cornea and the crystalline lens. The degree of eye elongation shows a strong negative correlation with the initial refractive error, suggesting an important visual effect in emmetropization [11, 12, 17].

Among the refractive components of the eye, the crystalline lens appears to play the most important role in refractive development beyond the age of 6-7 years [18], while the cornea has a smaller role [11]. The contribution of the cornea to emmetropization is inferior to that of lens because the cornea's development is virtually complete by the age of 2 years and, thus, it is unlikely to play an active role in maintaining emmetropia during ocular growth in childhood [18]. Several longitudinal and cross-sectional studies have shown that corneal power alters little across the school years [16, 19, 20]. The crystalline lens, on the other hand, tends to flatten, thin and lose power in this period of development [20-22].

Myopia is mostly due to excessive elongation of the axial length (particularly the vitreous chamber depth [23, 24]), while other dimensions of ocular size remain approximately constant. Elongation of the eye by merely 1 mm without other compensation will result in myopia of -2.0 to -2.5 Dioptres (D) [24]. The two most striking differences in ocular component changes between children with persistent emmetropia and those who develop myopia is the axial length and vitreous chamber depth: myopes show a lack of slowing in their growth [25].

In some cases, the cornea may undergo a "paradoxal" steeping during the "slow" phase of axial elongation and lead to myopia acceleration. Thus, myopic eyes usually have greater mean corneal power than emmetropic eyes [16].

The mechanism of emmetropization is poorly understood. However, it is known that a vision-dependent mechanism guides refractive state towards emmetropia [11, 13, 23]. The requirement for vision for eye growth regulation was proven in animal experiments: various species are known to develop ametropia when deprived of form vision or of clearly focused vision [26, 27]. In addition, a nonvisual mechanism operates without the need for visual guidance with the eye approaching emmetropia simply as a result of increasing eye size [11, 28].

1.1.2 Aetiology of Myopia

The aetiology of myopia is multifactorial, meaning that both genes and the environment play important roles. Twin studies indicate a strong genetic influence and a weak environmental impact, while differing myopia prevalences in different population groups from the same gene pool point to the opposite [29].

1.1.2.1 Genetic Factors

There are several signs of genetic influence on myopia development. Firstly, Myopia shows high heritability in twin studies (Table 1.1). Furthermore, ocular component dimensions exhibit high heritability as well: axial length 59-92% [30-32], lens thickness 93% [30] and corneal curvature 50-90% [30-32]. There also is a significant effect of the number of myopic parents on the risk of developing high myopia (odds ratio > 5.5 (95% CI: 3.2-12.6) if at least one parent is highly myopic) as well as strong association between axial length and parental myopic state [33].

Table 1.1 Recent Heritability in the Studies of Myopia (twin studies)

Heritability of myopia	Source
84 %	Hammond et al [34]
94 %	Lyhne et al [30]
75-88 %	Dirani et al [31]

The genetic component of myopia is confirmed by the findings that the prevalence of shortsightedness in certain ethnic groups sharing the same environment is different: (1) in the USA: Asian and Jews have high, while Africans and African-Americans have a low myopia rate [35]; (2) in Hawaii: Chinese have greater prevalence of myopia than Koreans, Japanese or Caucasians [36] and (3) in Taiwan: the frequency of myopia among purely aboriginal children is smaller than that in Chinese children [37].

Finally, the genetic background of myopia is also supported by the number of successful segregation and candidate gene analyses (section 1.4).

1.1.2.2 Environmental Factors

Evidence of an increase in the prevalence of myopia brings attention to environmental factors in the aetiology of myopia [38-40].

One of the environmental effects that has been much discussed is near work. The strong correlation between education and myopia [38, 41, 42] supports the idea that excessive accommodation produces myopia. A higher risk of myopia is observed in students when they are engaged in excessive near work (University term time), compared to when taking summer or winter vacations (breaks from near work) [43]. In addition, an interesting study conducted in Israel found that boys studying in orthodox school have higher myopia prevalence as compared to boys of identical ethnic background studying in general school. The authors attribute this difference to unique study habits of orthodox school and to the fact that the printed letters in the commentaries studied may be as small as 1mm in height [44].

However, attempts to reduce myopia progression with reading glasses or contact lenses have been disappointing [45]. Thus, near work is likely to be a weak risk factor for myopia, or else these interventions are not having the expected effects.

Moderate to severe myopia can be induced by optical alterations during emmetropization in the developing eye [6]. Visual impact on emmetropia development has been proved in several animal studies [11, 13, 23]. However, the applicability of these studies to human myopia is uncertain: for example, patients with form

deprivation, such as unilateral ptosis or congenital cataract, do not always develop myopia [46].

Apart from near work, several other environmental factors – such as education, diet/nutrition, psychology/personality, season of birth, maternal age and birth order, premature birth, low birth-weight and outdoors activities - have been proposed to be important in myopia development. These factors are summarized in Table 1.2.

In conclusion, myopia is a multifactorial disorder as both environmental factors along with genetics lead to its development.

Table 1.2 Environmental Aetiology of Myopia

Environmental Factor	**Support**	**Reference**
Near work	The greater the amount of near work, the higher the myopia degree or the more it progresses	[43, 44, 47, 48]
Education	Linear correlation between education and myopia	[35, 38, 41, 42]
Diet/Nutrition	Role of vitamins and minerals in growth and development	[49, 50]
Season of Birth	Significant association between season of birth and high myopia	[51, 52]
Maternal Age and Birth Order	Significant association of reduced vision with these factors	[53]
Premature Birth	Prematurely borns have a higher risk of developing myopia in latter life	[54]
Low Birth-Weight	Low birth-weight presents a risk factor for myopia development in latter life	[55]
Family Income	The prevalence of myopia increases as family income rises	[35]
Out-door activities	Out-door activities reduce the prevalence of myopia	[56]

1.1.3 High Myopia as a Significant Problem

High, or pathologic, myopia usually refers to refractive error worse than -6.0 Dioptres. This form of myopia is a health issue not only due to the need of glasses or contact lenses, but also because of its association with high level of ocular morbidity.

Pathological myopia carries an increased risk of additional eye disorders (Table 1.3). Some of these changes occur in the myopic eye only, whereas others can occur regardless of refractive error but have a greater prevalence in myopic eyes [57]. Pathologic transformations of posterior pole can reduce central vision to blindness. In addition, changes in retinal periphery are an even greater threat to vision because of the possibility of retinal detachment, resulting in complete loss of vision [57].

High myopia complications are recognized as a significant cause of visual impairment, especially because myopia-related blindness often affects people earlier in life when they still can be active professionally [24].

Table 1.3 Pathologic Changes in Myopic Eye
(Taken from Grosvenor and Goss [57])

Anterior Fundus	
Optic Nerve Crescents	Due to pulling away the choroid and pigment epithelium from the optic nerve head, allowing scleral tissue to be seen
Posterior Staphyloma Formation	Outward bulging of the eye over a restricted area due to localized weakness of the underlying sclera
Retinal Hemorrhages	Small, round hemorrhages near the macular area; a variable degree of vision loss may occur
Subretinal Neovascularization	Neovascular membrane is formed beneath the retinal pigment epithelium; the newly formed vessels are prone to leak
Posterior Fundus	
Vitreous Detachment	Characteristically occurs at the optic nerve head and can lead to retinal detachment
Retinal detachment	Sensory retina separates from the pigment epithelium; due to retinal breaks or tearing that are common in myopic eye
Fuchs' Spot	A lesion in macular (paramacular) area; is due to breaks in Bruch's membrane; may be cause of loss of central vision

1.2 Genetics

1.2.1 The Human Genome at a Glance

1.2.1.1 Chromatin

Nuclear DNA is packaged into a complex referred to as chromatin (with proteins called histones) [58]. The DNA is wound around a core of basic histones to form a structural unit called the nucleosome [59]. Chromatin structure is dynamic, accommodating the need for DNA to participate in various functions that require it as a template [60].

Traditionally, chromatin has been divided into hetero- and euchromatin depending on its accessibility for transcription. However, the genome is now known to be modified during gene expression to a higher degree than was previously anticipated (see below). In addition, transcription is now known to arise from intergenic regions, intron sequences and other non-coding genetic regions [61].

1.2.1.1.1 Euchromatin

Euchromatin represents areas of "active" chromosomal DNA available for transcription. A gene can be considered as a functional unit encoded in the genome, transcripts of which can be used directly (e.g regulatory RNA) or be interpreted in peptide (e.g. messenger RNA) [62].

Each gene has a specific position on a chromosome called its locus. Most genes consist of coding (exon) and non-coding (intron) regions. At the junctions between introns and exons there are highly conserved sequences (e.g. at the 5' and 3' ends of the inrons GT and AC dinucleotides occur), that are critical for normal splicing of messenger RNA (mRNA) [63].

At the 5' flanking region of the gene there are typically 3 "boxes" of homology: the CACCC box, the CCAAT box and the TATA box. All three boxes are conserved sequences and are generally required for accurate and efficient initiation of transcription - that is they are the major promoter regions for structural genes. At the 3'

non-coding region a polyadenylation signal (e.g. AATAAA) serves to recruit the machinery for end processing and polyadenylation of the 3' end of mRNA. [63].

Exons constitute three regions within the gene: (1) a region for RNA transcription; (2) a region translated to amino acid sequence; and (3) a region for the termination of translation [64].

Introns account for at least 30% of the human genome and may be a significant source of regulatory RNA [65]. Although the role of introns is still far from clear, non-coding RNA may determine many of our complex characteristics, play an important role in disease and contribute to genetic variation [66].

1.2.1.1.2 Heterochromatin

Heterochromatin refers to transcriptionally "inactive" stretches of DNA. Some areas of heterochromatin remain condensed throughout the organism's lifetime (constitutive or permanent heterochromatin), while others can be assembled when needed (facultative or optional heterochromatin) [67].

Regions important for the genome integrity (e.g. bands of satellites present next to centromeres [67]) , repetitive and noncoding sequences, are kept stably as constitutive chromatin [68]. Despite their condensed state, transcription from these regions is possible. However, transcript levels are low and do not match those of euchromatin [69].

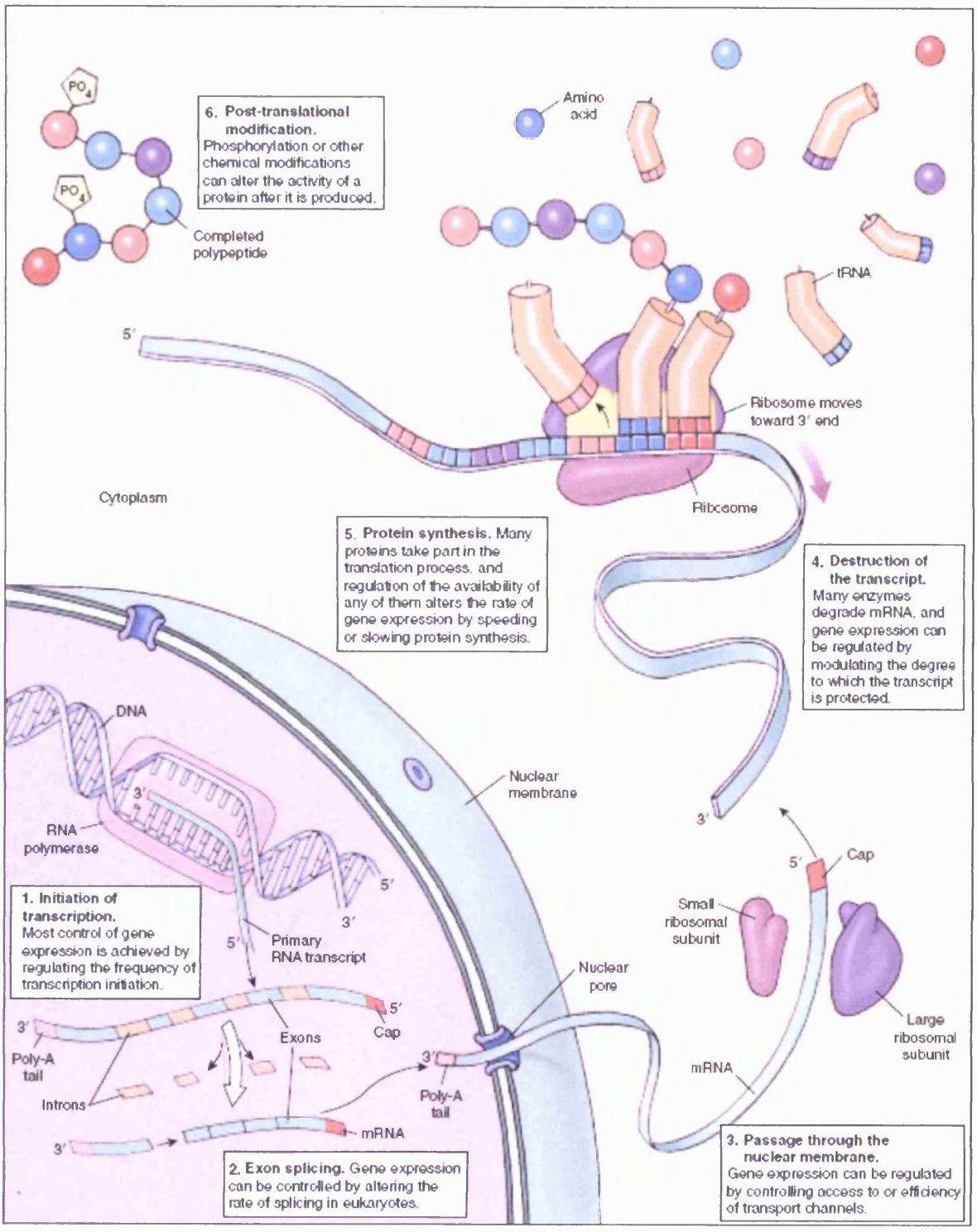
Similarly, permanently condensed chromatin, facultative chromatin is also inert to transcription, but maintains the potential to convert to euchromatin [69]. The major components of heterochromatin are the regulatory factors of DNA and histone methylation, transcriptional repressors and functional RNA [67].

1.2.1.2 Gene Expression and its Control

The general concept of gene expression is that of a pathway from DNA to a polypeptide via chromosome structure, primary transcript, mRNA processing and translation [70]. However, genomic expression is not as tightly related to protein formation as had been thought [71]. The classical model of most genetic information being translated into proteins is now challenged by recent evidence suggesting that the majority of the genome is transcribed into so-called non-coding RNA (ncRNA) [66].

The control of gene expression can occur on 2 major levels: transcriptional and post-transcriptional regulation [72] (Figure 1.5). The majority of regulatory events, however, happen at transcription [62].

Figure 1.5 Steps of Gene Expression Control in Eukaryotes
(Taken from Raven &Johnson [72]).



Firstly, for transcription initiation, DNA needs to be accessible: a gene may exist in either of two structural conditions and, thus, the change between permissive and non-permissive chromatin states leads to activation or repression of transcription. The alteration in structure is associated with histone acetylation and gene methylation [59]. The alteration in structure is associated with so-called epigenic control (section 1.3.1.3).

Once nucleosomic DNA is made accessible, transcription can start. This requires a close collaboration of transcription factors (*trans*-acting proteins) and regulatory DNA sequences (*cis*-acting modules).

Trans-acting transcription factors (TF) assemble into a complex with RNA polymerase and cis-acting DNA sequences in such a way that transcription can be initiated and tightly controlled at the same time. These factors can be divided into 2 groups: (1) constitutively active, basal factors which stabilize and guide RNA polymerase binding to a promoter; (2) regulatory - so-called coactivators, enhancers and repressors – factors that interact with regulatory DNA modules and other proteins [73]. The first group are essential for transcription, but cannot by themselves increase or decrease its rate. The latter, on the other hand, positively or negatively affect the pace of transcription by binding to governing cis-elements [72].

C*is*-acting expression-control DNA sequences may be located within genes or in intergenic regions [74]. These modules can also be divided into 2 groups: (1) promoters; (2) transcription rate controlling enhancers and silencers [73]. In contrast to promoters, the positions of the latter sequences are variable with regards to the genes they are regulating. Located upstream or downstream of a gene, enhancers activate transcription in a distance- and gene-independent manner. Silencers, on the other hand, increase the probability that a gene is repressed in any given cell [75].

Post-transcriptional control of gene expression can occur via the processing of a primary transcript (RNA splicing), selective degradation of mRNA or translation rate control [72]. An example of such regulation is the occurrence of a premature termination codon that precludes the synthesis of a full-length protein, resulting in a non-functional, truncated gene product. Approximately 33% of inherited and acquired

Mendelian diseases are attributable to a premature termination codon [76].

The expression level of many genes shows natural variation, which is probably due to polymorphisms in DNA sequence. This variation is likely to account for a substantial part of human diversity and has a heritable component [74]. It can, therefore, contribute to differences that are important for understanding the aspects of the susceptibility to complex diseases [77].

1.2.1.3 The Human Epigenome

The epigenome refers to chemical modifications of DNA bases and histone proteins, forming a complex regulatory network that modulates chromatin structure and genome function, influencing how the genome is made manifest across a diverse array of developmental stages, tissue types and disease states. Although these chemical changes and are not encoded in the nucleotide sequence, they are potentially heritable [78].

Epigenic modifications fall into two main categories: DNA methylation and histone modification. In humans, DNA methylation occurs almost exclusively in the context of so-called CpG islands [82, 83]: regions of DNA with a high G and C content and high frequency of CpG (phosphodiester bonded C and G nucleotides) dinucleotides relative to the rest of the genome [79]. Such islands cover about 0.7% of the human genome [80] and are associated with about 60% of human gene promoters [78]. A methylated cytosine base can promote or preclude the recruitment of gene expression regulatory proteins through methyl-CpG binding proteins. The preservation (or inheritance) of methylation is thought to be mediated by a methyl-transferase enzyme, which has specificity for hemi-methylated CpG dinucleotides: the enzyme methylates a newly synthesized DNA strand based on the presence of methylation in the CpG dinucleotide in the complementary template strand [82, 83].

The core histones are subject to more than 100 different post-translational modifications, including acetylation, methylation, phosphorylation and ubiquitination [78]. The vast majority of these modifications (including their inheritance), however, is still poorly understood.

DNA methylation has been implicated in a number of such cellular functions and pathologies as tissue-specific gene expression, cell differentiation, genomic imprinting, regulation of chromatin structure, carciogenesis and aging [81]. In cancer development, for example, characteristic epigenetic changes include hypermethylation of the tumor suppressor genes' promoters, which typically results in their silencing [82].

Another example of epigenetic control is imprinting: the phenomena of gene expression being dependant on the sex of the parent from whom the gene was inherited (parent-of-origin effect). Imprinted loci are characterized by the reduced or absent expression of either the paternally- or maternally-derived allele [83]. Approximately 1% of all human genes are thought to be imprinted [84].

Imprinting syndromes are a group of medical conditions that result from the altered expression of genes. These alterations, however, can be derived not only from epigenic control, but also from such changes in DNA sequence as (1) large deletions or duplications of chromosomal regions that contain imprinted genes; (2) DNA mutations and (3) uniparental disomy [85]. One of the most well observed imprinting syndromes is Prader-Willi syndrome (PWS), which results from the absence of paternal expression of a cluster of non-coding RNAs [85]. The underlying molecular mechanism in the great majority of PWS patients is either a 4-6 Mb chromosome deletion at 15q (70%) region or maternal disomy of chromosome 15 (25%). The rest (5%) may be accounted for by epigenetics in the form of hypomethylation of the paternally inherited allele [85].

1.2.1.4 Heterogeneity of The Human Genome

Each copy of the human genome is unique and differs in sequence from any other copy in the population. Despite the fact that 99.9% of the DNA sequence in two randomly selected individuals is identical, the variability in remaining 0.1% of DNA sequence is enough to influence human diversity in physical appearance, susceptibility to a disease or response to a medical treatment [86]. Although the relative contribution to complex human traits of DNA variants that alter protein structure, versus variants that alter the pattern of gene expression, is unknown [87], what is certain is that all variable human

traits are likely to have at least some genetic contribution [88].

The diversity in nucleotide sequence can occur in intergenic regions as well as within genes. Given the diploidy of humans, each locus is represented by 2 alleles (genotype) that can either be the same (homozygosity) or differ (heterozygosity). Polymorphic alleles that co-occur on a chromosome are called haplotypes.

In studying human genetic variation in its totality, it is crucial to sample subjects of diverse ethnogeography, as chromosomes sampled from different populations have substantial differences [89]. Performing genetic association studies, on the other hand, typically utilise participants of the same ethnicity to avoid spurious results due to allele frequency diversity between populations (section 1.3.4.2.3).

DNA sequencing and analysis has revealed several types of variability in the human genome: microsatellites, single nucleotide polymorphisms and structural variants, each of which is detailed below.

1.2.1.4.1 Microsatellites

Microsatellites, or tandem repeat loci, are characterized by numerous contiguous repeats of the same short sequence unit, typically ranging from 1 to 6 nucleotides in size. At these sites, the number of repeated copies varies greatly: many microsatellites have 5-10 alleles [90].

Approximately 3% of the human genome is occupied by microsatellites [91]. Detailed examination of repeat loci has revealed that mononucleotide repetitions are the most abundant class of microsatellites, while trinucleotide alleles are about three times less frequent than di- and tetranucleotide repeats [92]. The distribution of microsatellites within the genome is not random: typically, longer alleles occur within non-coding regions [93, 94].

Microsatellites located in promoter regions, untranslated regions and introns can be important regulators of such aspects of gene expression as translation rate, RNA stability and splicing efficiency [95, 96]. Microsatellites occurring in intergenic

regions may also have functional role: it has been suggested that tandem repeats can alter chromatin organization and may be associated with recombination hotspots [94].

Microsatelittes have been used for disease gene mapping since the late 1980s [97] as their alleles can be easily and rapidly distinguished on the basis of variations in electrophoretic movement of fluorescent-labelled PCR products [90].

Because microsatellites are highly polymorphic and contribute to gene regulation, observation of changes in the length of their alleles may provide a large pool of heritable variance. Indeed, allele length polymorphisms of microsatelites are implicated as genetic risk factors for such complex diseases as cystic fibrosis [98, 99] and breast cancer [100].

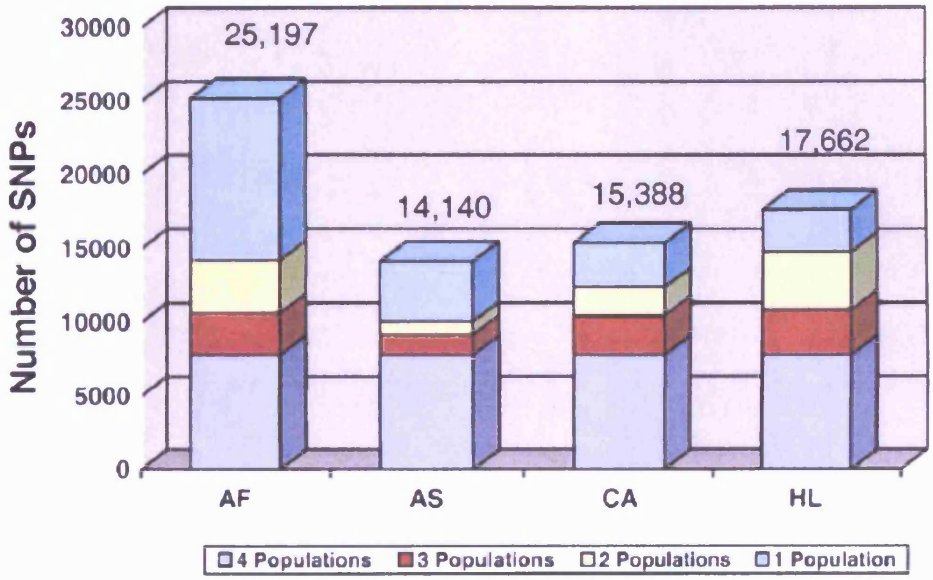
1.2.1.4.2 Single Nucleotide Polymorphism

Single nucleotide polymorphisms (SNPs) are the substitution of a single base and are the most common form of DNA variation [87]: there are about 15 million SNPs in the human genome [89].

The presence or absence, as well as the frequencies, of SNPs vary considerably among gene regions and among populations. A number of population-specific SNPs with minor allele frequency (MAF) substantially above 5% have been observed in one population but not in another, demonstrating an appreciable variation in SNP frequencies among human populations (Figure 1.6) [88]. The largest number of population-specific SNPs has been found in individuals of African origin [88, 89, 91].

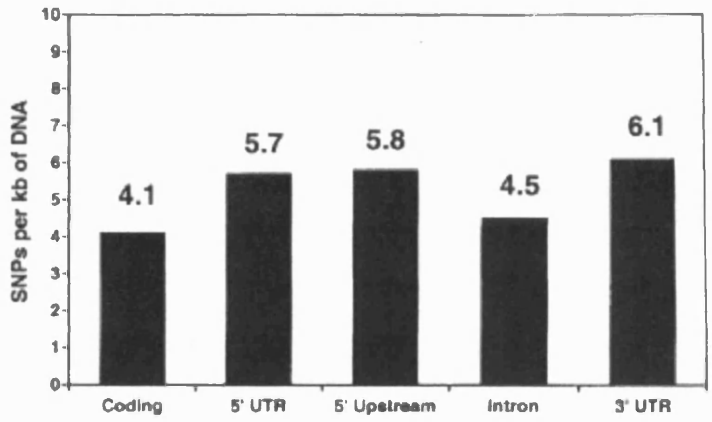
Figure 1.6 Population Distribution of 37582 SNPs among Individuals of Different Ethnic Origin

(AF: African-American; AS: Asian; CA: Caucasian; HL: Hispanic-Latino). The degree of population sharing is indicated in color. Over 2/3 of the polymorphisms observed were variable among AF individuals, whereas between 37 and 47% of the SNPs were variable in each of the other populations (Taken from Schneider et al. [88]).



Regions of DNA that affect gene expression are highly variable, containing 0.6% polymorphic sites [89]. The distribution of SNPs in various genomic regions suggests that there is conservation of the coding region (Figure 1.7): the average gene contains about four SNPs in its coding sequence, with allele frequencies of at least a few percent [101]. However, the SNP density varies less than 2-fold among all regions and could be even higher in large introns or in intergenic regions [88].

Figure 1.7 SNP Distribution per kilobasepair of Functionally-defined Genomic Regions of 1630 genes
(Taken from Schneider et al. [88])

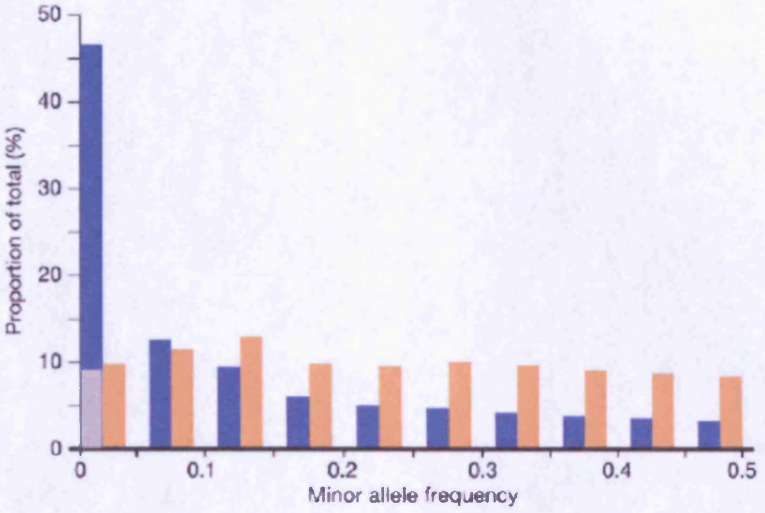


A large proportion of SNPs in the human genome show a minor allele frequency of less than 1% (Figure 1.8): most SNPs observed in ENCODE regions have MAF < 5% and 9% of them were only observed in a single individual (Figure 1.8). Although the majority of polymorphisms in the population are rare, most heterozygous loci within any individual are due to common SNPs [102] and it is frequently suggested that SNPs whose MAF > 5% are of most interest in disease gene mapping (see below) [101].

SNPs represent a great interest in human genetics. Found in a functional gene region, a SNP may encode a difference in protein sequence or expression, which in turn can lead to a disease or other phenotype. It may also mark the presence of other, perhaps less easily detected, sources of genetic diversity that cause a phenotype of interest [101].

Figure 1.8 Minor Allele Frequency Distribution of SNPs in the ENCODE Regions

Polymorphic SNPs are shown according to their MAF (blue). The sum of contribution of each MAF bin to the overall heterozygosity is shown in orange. SNPs that are heterozygous in one individual only are marked as grey (Taken from International HapMap Consortium [102]).



1.2.1.4.3 Structural Variants

Structural variants are defined as genomic alterations that involve segments of DNA of 1000 bases (1 kb) or larger [103].

In contrast to SNPs that affect only a single nucleotide, structural variants can affect from one kilobase to several megabases of DNA per event (deletion, insertion, duplication and complex rearrangements of genomic regions), adding up to a significant effect on phenotypic variability. Table 1.4 comprises a brief description of commonly encountered types of structural variants.

The discovery of structural variants demonstrates the plasticity of the human genome and might help to explain phenotypic discrepancies in genetic traits and/or in the severity of the resulting phenotype. It may also provide new leads for the molecular basis of complex disorders [104].

Table 1.4 Structural Variation Definitions
(Adopted from [103] and [104])

Structural Variant	Definition
Copy-number Variant (CNV) or Polymorphism	A segment of DNA that is 1kb or larger and is present at a variable copy number in a specific chromosomal region. If its population allele frequency is less than 1%, it is referred to as a rare variant; if its frequency exceeds 1%, the term copy number polymorphism may be used.
Segmental duplication or low-copy repeat	A segment of DNA > 1kb in size that occurs in two or more copies per haploid genome and the different copies share >90% sequence identity. They are often variable in copy number and can therefore also be CNVs.
Inversion	A segment of DNA that is reversed in orientation with respect to the rest of the chromosome and to a specific reference genome.
Translocation	A change in position of a chromosomal segment within a genome that involves no change in the total DNA content. Translocation can be intra- or inter-chromosomal.

1.2.1.5 Utilization of Genetic Variation in Genetic Studies

The ability to genetically map complex disorders is facilitated by identifying and genotyping DNA polymorphisms termed "markers". These DNA variations used in genetic analyses can be divided into 5 groups: (1) restriction fragment length polymorphisms (RFLP); (2) variable number of tandem repeats (VNTR); (3) microsatellite or short tandem repeats; (4) SNP and (5) copy number variations (CNV). The type of polymorphism utilized in human genetics has been changing with time.

Until recently, analyses have been based on widely spaced (usually ~10 cM) microsatellite markers, but it is now possible to genotype a dense map of SNP markers at low cost. In addition, construction of the international SNP database (HapMap) enables the performance of not only genome-wide analyses, but also of candidate gene studies facilitating the choice of so-called tagging SNPs (section 1.3.3.2). Apart from SNPs, copy number variations are of particular interest in current human genetics. However, development of novel techniques and statistical methods is needed to capture this new form of genetic variation in a meaningful manner [105].

There has been a debate whether it is better to use microsatellites or SNPs in genetic studies. Although SNPs are somewhat less informative than microsatellites, the current trend is to use single nucleotide polymorphisms: it is technically easier and less expensive to genotype SNPs because they have only 2 alleles and require less DNA [106]. Furthermore, microsatellites have a disadvantage of being prone to mutation, which makes their use more challenging compared to SNPs. It is noteworthy, however, that because of large genetic variability of microsatellites, the chance of finding disease causative allele in linkage disequilibrium (section 1.3.3.2) with such marker is much higher than with SNP markers.

1.2.2 Phenotype and its Inheritance

1.2.2.1 Phenotype as the Result of a Genotype

A phenotype is the observable expression of an individual's genotype [107]. While genotypes act through proteins and different molecular pathways remaining mostly stable over the lifetime of an individual, phenotypes are observed through signs, symptoms and visible traits and are often dynamic. Therefore, genetic studies need to pay particular attention to issues of phenotype definition and measurement: the phenotyping needs to be standardized to increase the quality of research and the reproducibility of linkage and association studies [108].

The presence (affected) or absence (non-affected) of a certain trait depends in part on an individual's nucleotide sequence: a genetic variant shared by all affected (but not by non-affected) subjects is likely to be responsible for the trait examined. When such a variant is found, the probability that a randomly selected individual who caries the variant will be affected can be estimated as the phenotype penetrance [109]: for example, the ratio of risks for developing a phenotype between those with and those without the susceptibility genetic factor (allele, genotype or haplotype). These ratios are often used as a criterion of association between a trait and a genetic variant (genotype relative risk).

A number of different factors (of both genetic and environmental origin) determine the relationship between genotype and phenotype: two individuals with identical genotype at a given locus can experience different clinical symptoms due to a differing genetic background [110]. Such elements include the pattern of inheritance, allelic heterogeneity, locus heterogeneity, variable penetrance, epistasis as well as environmental variables (Table 1.5).

Table 1.5 Sources of Heterogeneity in Susceptibility to Complex Diseases [111], [112]

Heterogeneity	Explanation
Locus Heterogeneity	Phenotypically indistinguishable diseases caused by mutation in one of two or more separately located genes.
Allelic Heterogeneity	Different mutations or deletions within a single gene may cause a common disease phenotype.
Epistasis (Gene Interaction)	The possession of a certain mutation or genotype will confer susceptibility to a degree dictated by the presence of other mutations or genotypes.
Environmental Vulnerability	Phenotypes are influenced by environmental stimuli.
Gene x Environment Interactions	Gene or genes have their effects only in the presence of particular environmental stimulus. Strictly, a genotype leads to a different phenotype depending on the environment in which it occurs.

1.2.2.2 Phenotype in a Genetic Study: Discrete and Continuous Traits

In genetic studies, traits are classified as discrete or continuous. The term discrete phenotype applies to those traits that are either present or absent: such as cancer or retinitis pigmentosa.

A continuous phenotype, on the other hand, has a range of possible values and these values are often used directly. Such quantitative traits include for example body weight and height, blood pressure and refractive error. Continuous traits can also be categorized as dichotomous by using a predefined threshold value; sometimes, especially for genetic studies, only individuals in the extremes of the frequency distribution are used in order to maximize power and obtain a definitive distinction between diseased and nondiseased individuals [107].

1.2.2.3 Inheritance of a Phenotype: Mendelian and Complex Traits

To be able to understand the spectrum of human genetic disease, it is essential to consider the way in which genes may be inherited. Some inherited disorders follow a simple Mendelian form of transmission. Complex or multifactorial traits, on the other hand, are determined by a number of genetic and environmental factors [111]. In contrast to monogenic Mendelian phenotypes that are controlled by single genes, complex traits are defined by multiple genes and are therefore called multigenic traits [113]. Many Mendelian phenotypes vary in diverse biological features such as age of onset or severity, suggesting that genetic background tends to modify the phenotypic expression leaving few if any Mendelian disorders to be truly monogenic. Common diseases are almost always genetically complex [114], since otherwise robust selection wold be expected to reduce the risk allele frequency in a population.

A phenotype is considered dominant if it appears in the heterozygote in whom only one allele is defective. Dominant mutations often result in a clinical symptom by giving rise to reduced or abnormal expression of a gene product. In a recessive disorder, both alleles must be mutant (homozygous state) for a phenotype to become apparent. A recessive allele does not necessarily lead to a disease trait: production of 50% of the normal level of the gene product in a heterozygote may be sufficient to avoid clinical symptoms [110].

Mendelian dominant or recessive inheritance can also be either autosomal or sex-linked, depending on which chromosome the mutant allele appears on. In case of an X-linked recessive trait, males are affected more commonly since they do not carry a homologous X-chromosome which can serve to mask clinical expression of the disorder in females. By contrast, in X-linked dominant disorders, the mutation is manifested more equally in females and males, although the absence of a normal allele often results in males being more severely affected than females [110].

1.2.2.4 Common Disease Traits: the Genetic Challenge

Identification of the genes that contribute to complex traits poses a special challenge because of their high genetic heterogeneity. To address this issue, two main hypotheses have emerged regarding the genetic susceptibility to common diseases: the common disease-causing variant hypothesis and the rare variant hypothesis [115].

The common disease-common variant (CDCV) hypothesis posits that a few common allelic variants (defined as having a MAF > 1% [116]) account for much of the genetic variance in disease susceptibility [117]. DNA variants leading to monogenic diseases are usually rare due to natural selection. By contrast, because variants in genes involved in polygenic traits do not act alone to produce the phenotype, selection against them will only occur when they are present in the disease-causing combination. Thus, these variants may exist at a high frequency in the population [118].

An alternative, although not mutually exclusive, hypothesis is that genetic susceptibility to common disorders is due to summation of the effects of a series of rare variants in different genes, each contributing a more substantial increase in relative risk [115]. Such rare variants will mostly be population specific because of founder effects resulting from genetic drift [119].

A critical feature shared by common and rare variants is that they do not necessarily give rise to a familial concentration of cases (as opposed to familial segregation of Mendelian traits). This is because the penetrance of such variants is low. Most of the common alleles found so far are associated with risk ratios of only between about 1.2 and 1.5, while rare variants, on average, show risk of 2 or more. Only when penetrances are well above 50% does one approach a familial concentration that begins to look like a standard Mendelian segregation [119]. Other, general and individual properties of these variants are listed in Table 1.6. Whether common disorders are primarily caused by common or rare variant is still an open question. The current literature suggests that both these hypotheses are correct, depending on the gene and disease examined.

Table 1.6 Characteristics of Common and Rare Variants
(Taken from Bodmer and Bonilla [119]).

	Common Disease Variants	**Rare Disease Variants**
Discovery	Population association studies	DNA sequencing of candidate genes
Minor Allele Frequency	> 5%	> 0.1% to 2-3%
Risk Ratio	1.2 – 1.5	> 2.0
Familial concentration of cases	None	None
Contribution to Disease Aetiology	Hard to find functionally relevant variant	Functionally relevant, often obvious variants

1.2.3 Independent Assortment and Linkage Disequilibrium

1.2.3.1 Independent Assortment of Gametes and Recombination

A diploid organism, such as human, produces a large number of genetically unique gametes. Cells undergo two processes to produce this diversity: independent assortment and recombination [120].

Independent assortment is the random distribution of maternal and paternal chromosomes into gametes during meiosis. Once the first gamete is drawn from an individual, the second one still has an equal chance that its chromosomes stem from either parent. In humans, independent assortment yields 2^{23} (over 8 million) unique ways to distribte the 23 pairs of chromosomes [120].

Recombination is the result of "crossing over", which occurs early in meiosis when the homologous chromosomes exchange DNA, such that each caries some paternally and some maternally derived genetic material (Figure 1.9) [120]. Genes on the same chromosome are physically linked and thus tend to be co-transmitted from one generation to the next, each serving as a marker for the other. However, meiotic recombination can lead to segregation of these alleles to different germ cells, so that an individual may inherit a new combination of alleles. When recombination fails to cause segregation of these alleles, they are said to be in linkage disequilibrium [118].

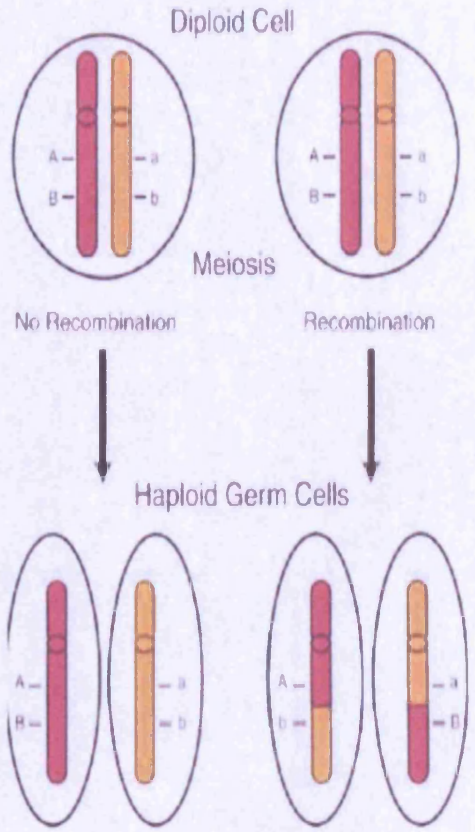


Figure 1.9 Meiotic Segregation of DNA Variants
(Taken from Schafer et al. [118])
A/a and B/b are alleles at two loci. "A" and "B" (as well as "a" and "b") are physically linked on two chromosomes. Recombination between chromosome pairs can result in chromosomes with A/b on one chromosome and B/b on another. When recombination fails to cause segregation of the alleles, they are said to be in linkage disequilibrium.

1.2.3.2 Linkage Disequilibrium and its Estimation

Linkage disequilibrium (LD) refers to the non-independence of alleles at two or more loci. When a sample of chromosomes is drawn from a population, all the chromosomes are related by some ancestral genealogy. Thus, genetic markers that are very close together on a chromosome have either the same or similar genealogies and this induces dependence between them. Markers that are further apart may have different ancestry because of recombination and, for this reason, the strength of LD between pairs of markers decreases as a function of genetic distance between them [121]. Nonetheless, local variation in LD overwhelms this "rule" over short distances: markers that are adjacent to each other on a chromosome may be statistically independent, whereas those that are further apart can be highly correlated [122].

With most human recombination occurring in recombination hotspots, the breakdown of LD is often discontinuous creating a "block-like" structure. However, the tendency towards co-localisation of recombination sites does not imply that all haplotypes break at each recombination hot spot [123].

Although genealogy and recombination provide insight into why nearby SNPs are often correlated, it is redundancy among genetic variants (e.g.SNPs) that are of central importance for the design and analysis of genetic studies. A truly comprehensive association study must consider all putative causal alleles and test each for its potential role in a disease. If a casual variant cannot be directly tested in the sampled population, its effect can nonetheless be examined indirectly if it is in strong linkage disequilibrium with a directly tested SNP. When two variants are perfectly correlated, testing one is exactly equivalent to testing the other. Thus, taking the number of distinct combinations of SNP alleles (haplotypes), it is possible to select a parsimonious set of SNPs that would capture the information of all variants that are in strong LD with these selected, so-called tagging SNPs and, therefore, distinguish the haplotypic variation in a population [124].

Various statistical measures can be used to assess LD between a pair of loci, but in practice only two, namely D' and r^2 are widely used.

In what follows, the discussion will be restricted to a marker and a disease locus each having two alleles: disease alleles "A" and "a"; and marker alleles "B" and "b". Thus, the haplotypes for the disease and the single marker can be arrayed in 2x2 table with marginal probabilities p_A, p_a, p_B and p_b for each allele of these 2 loci (Table 1.7).

In principal, D' and r^2 measures of LD reflect the difference between the observed and the expected (under independence) frequencies of haplotypes bearing the disease and normal alleles: $D = p_{BA} - (p_B \times p_A)$ [125]. The so-called D' measure can be obtained by normalizing this D value by the absolute maximum D that could be achieved given the table margins [126]; while raising D to the power of 2 and dividing it by the multiple of all marginal frequencies will result in r^2 ($r^2 = D^2/p_A \times p_a \times p_B \times p_b$) [121]. Both D' and r^2 have the same scale from zero to one: zero implies independence and one means complete LD between the two loci [122].

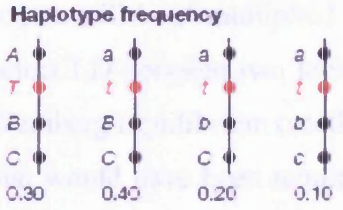
Although the mathematical interpretation of these measures may seem to be the same, it is important to understand the difference in their practical meanings: r^2 equals one only when the two loci have identical allele frequencies and every occurrence of an allele at each of the markers perfectly predicts the allele at the other locus. By contrast, D' can reach a value of one when the allele frequencies vary, as it reflects the correlation only since the most recent mutation [122]. Thus, D' can be large even when one of the alleles is very rare (Figure 1.10), which is usually of little practical interest in disease gene mapping [127].

Table 1.7 Layout and Notations for Sample Haplotype Frequencies.

(p denotes the frequencies and marginal probabilities of the haplotypes in the sample)

	Disease Allele "A"	**Disease Allele "a"**	**Total**
Normal Allele "B"	p_{BA}	p_{Ba}	p_B
Normal Allele "b"	p_{bA}	p_{ba}	p_b
Total	p_A	p_a	1

Figure 1.10 A Hypothetical Allele Frequencies and Haplotype Set of Different Situations of LD Between a Disease and Marker Loci
(Adopted from Zondervan and Cardon [122])



Allele frequency	D	$D'_{(marker, T)}$	$r^2_{(marker, T)}$
A = 0.30	0.21	1.0	1.0
T = 0.30			
B = 0.70	0.09	1.0	0.18
C = 0.90	0.03	1.0	0.05

Consider a diallelic locus in which allele "T" is associated with a complex disease. The population frequency of "T" is 0.3 (thus, the frequency of "t" will be 1-0.3 = 0.7). Cases and controls are collected for the study of the disease, but the disease polymorphism is not genotyped. Instead, three surrounding SNPs (A/a, B/b and C/c) with different allele frequencies and different LD relationships with the disease allele are typed. Note that the haplotypes carrying the "T" allele will be over-sampled in cases relative to its frequency in the population as a whole. The trait allele "T" is present only in one haplotype, ABC, which has a frequency of 0.3 in the population. Marker allele "A" is also present only on ABC, and therefore has the frequency of 0.3. Allele "B" occurs on ABC and on aBC, with a total frequency of 0.7, whereas "C" occurs on ABC, aBC and abC with a total frequency of 0.9. All three marker alleles "A", "B" and "C" are in complete LD with the disease allele "T" in terms of D', but not all B or C alleles are co-inherited with "T". Thus, marker allele "A" is the only one with an r^2 value of one.

Allele frequencies of marker B/b are identical to those of marker A/a, as well as for the trait locus. The D' values between "B" and "T" and between "A" and "T" are both one. However, the "B" allele is not the one that matches the disease allele frequency: it is the frequency of the other allele "b" that does, but it never occurs on the haplotype with the disease allele. For equal statistical power, it would take a sample size 5.5 times greater to detect a disease association with "B" than with "A" ($1/r^2 = 1/0.18 = 5.5$). This shows that markers with equal MAF and high D' are not sufficient to ensure high power; they must match in phase as well [122].

In Table 1.7, the observed proportions of gametes in a population sample of size N, the χ^2 test for association between the loci would be: $D^2N/p_A \times p_a \times p_B \times p_b$ [128]. Replacing $D^2/p_A \times p_a \times p_B \times p_b$ with r^2, the test statistic for independence of haplotype counts will be r^2 multiplied by the sample size. Consequently, r^2 reflects the power to detect LD between two loci. If disease risk is multiplicative across alleles and Hardy Weinberg Equilibrium (section 1.3.3.3) holds, the reciprocal of r^2 gives the sample size that would have been required to detect the disease association by directly typing the casual polymorphism, relative to the sample size required to achieve the same power when typing the marker (Figure 1.10) [127].

The performance of D' and r^2 greatly depends on variation in marker allele frequencies and on the configuration of markers surrounding the disease locus (Figure 1.10). The value of D', for example, is independent from the marginal allele frequencies in mathematical terms, but it is not in any other general sense (e.g. the force of equal recombination rate on different populations with different allele frequencies will result in unequal D' values for these populations) [128].

The correlation between a causal mutation and haplotype on which it arose – linkage disequilibrium – has great value for both fine-mapping and genome-wide genetic association studies. However, the actual degree of disequilibrium between two loci is drawn from a probability distribution that results from the evolutionary process: LD can be influenced by other phenomena besides recombination, namely mutation, drift, mating choice and selection. These population genetic phenomena can mask the impact of recombination, leading to a large variance in LD values [125].

1.2.3.3 Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium implies constant genotype frequencies from generation to generation in a population whose members are mating randomly, with no selection or migration. Under such an equilibrium, the genotypic frequencies at an autosomal locus with two alleles ("A" with relative frequency "p" and "a" with relative frequency "q") will be expected to be p^2 for genotype "AA", 2pq for genotype "Aa" and q^2 for genotype "aa". In addition, all three genotypic proportion will sum to one, as will the allele frequencies [129]. HWE, thus, depends on a series of assumptions about the

tested population, including, for example, that no new mutations arise, no selection occurs and mating is random [130]. Departures from HWE, if not due to a change or violation of these assumptions, may therefore point to genotyping error, population admixture or a true non-independence of alleles in the population (e.g. due to the influence of an allele on disease prevalence) [131, 132].

The most common two ways of assessing HWE are through a goodness-of-fit chi-square test and an exact Fisher's test (section 2.4.2.3). The performance of both tests depends on the sample size and minor allele counts. However, χ^2 tests tend to overestimate the significance level, especially in smaller samples, while the exact statistic never exceeds the nominal significance level [133, 134].

1.2.4 Genetic association

1.2.4.1 Transmission Disequilibrium and the Concept of Genetic Association Studies

Under the law of Mendelian assortment, alleles at a locus will be transmitted randomly and with equal probability from parents to an offspring. Deviation from the random occurrence of an allele regarding disease phenotype (transmission disequilibrium) is considered to be genetic association. Allelic association reflects sharing of ancestral chromosomes: alleles at loci tightly linked to disease susceptibility locus will be shared among affected individuals more often than expected by chance.

Classically, association can be examined with the transmission disequilibrium test (TDT), which compares the observed number of alleles transmitted to affected offspring with those expected in Mendelian transmissions in terms of chi-square statistics [135]. Originally, TDT was used to test for linkage in the presence of association. However, because its null hypothesis assumes both no linkage and no association, the TDT is now typically used as a test for association.

With time, several types of association studies have been developed. The next section of this chapter summarizes the current methods according to the markers utilized

(direct and indirect studies), aim (hypothesis generating and hypothesis testing studies) and the nature of examined cohort (family-based and case/control studies).

1.2.4.2 Testing for Genetic Association

1.2.4.2.1 Direct and Indirect Association Approaches

Most association studies rest on the assumption that linkage disequilibrium exists and, thus, the causal variant can be examined either directly (direct association) or by the means of a polymorphism in LD with it (indirect association). Commonly, the casual variant will not be typed in the study. Nonetheless, a well-designed experiment will have a good chance of including one or more genotyped polymorphisms that are in strong LD with a common casual variant and be able to detect the indirect association between marker locus and disease phenotype [127].

The limitation of association studies being indirect can be overcome by exploiting the block-like structure of LD, characterized by the existence of genomic regions with little evidence for historical recombination and limited haplotype diversity. Within such regions, genotypes of common SNPs can be inferred from only a few so-called tag SNPs (sections 1.3.3.2 and 1.3.4.3.2) [136-138]. Moreover, because LD is a short range phenomena, if association exists, it will define a small candidate region in which to search for a susceptibility gene.

1.2.4.2.2 Hypothesis-generating and Hypothesis-testing Association Studies

Association analyses can be used for the genome-wide, genetic exploration of a disease (hypothesis-generating) or for the identification of candidate polymorphisms (hypothesis-testing) (Table 1.8). This classification, however, is not precise: some candidate gene studies may involve many genes and are similar to genome-wide scans [127].

Genome-wide, exploratory, hypothesis-generating analyses present an opportunity to identify associations between genetic polymorphisms and a complex trait. For this kind of test, a large number of SNPs is typed throughout the genome (a high SNP density is

essential for mapping a putative association region). High-throughput genotyping makes this a realistic and affordable strategy.

Once the region of interest is known, the next step is to test the hypothesis and fine map the exact polymorphism responsible for a disease-related phenotype. In this case, markers can span a gene (candidate gene) or a locus on a chromosome (e.g. a linkage peak).

Table 1.8 Types of Population Association Studies
(Taken from Lewontin [127])

Type of an Association Study	Description
Candidate Polymorphism Study	Focuses on an individual polymorphism that is suspected to be the causal one.
Candidate Gene Study	Focuses on candidate gene (or genes) and involves the genotyping of several polymorphisms within that gene (or genes).
Fine Mapping	Focuses on a candidate region that has been identified by previous studies; may involve several genes with genotyping of hundreds of polymorphisms.
Genome-Wide Study	Focuses on identifying common casual variants throughout the genome.

1.2.4.2.3 Case/control and Family-based Analyses

Genetic association for complex traits can be assessed either with a case/control study of unrelated people or with a family-based design. Although these are two fundamentally different approaches, which have their own strengthes and weaknesses, these analysis should be viewed as complementary and not competitive in the effort to overcome the challenges of association studies [139]. Thus, combined case/control and family association studies can also be performed.

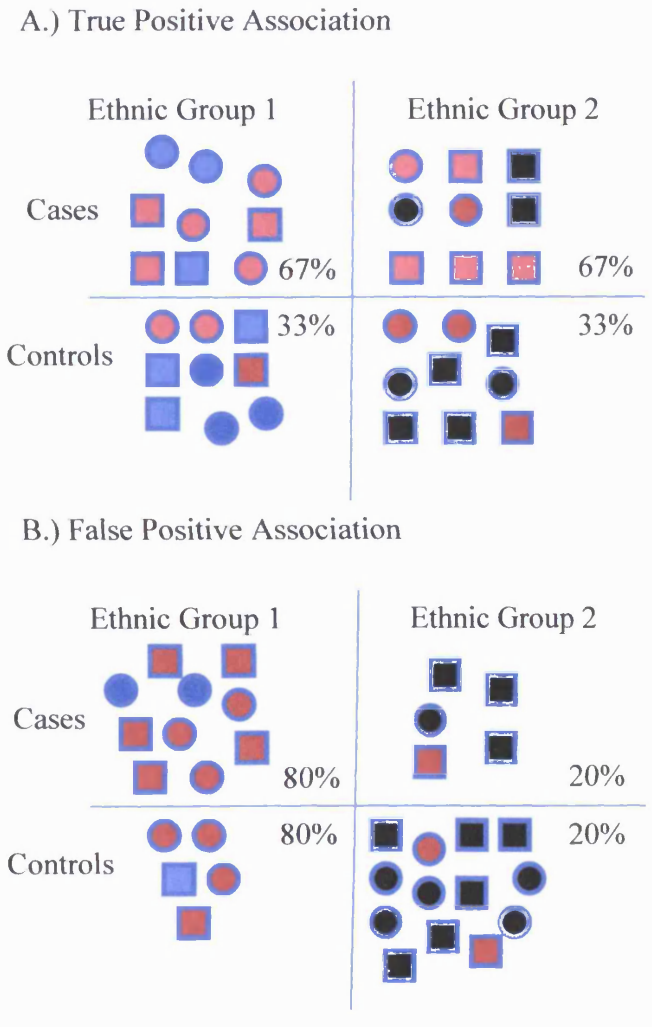
The classic case/control design compares allele frequencies of genotypes in a sample of unrelated affected and a sample of unrelated unaffected individuals [140]. The major criticism of these studies is the potential for spurious association due to population stratification: the existence of genetically different groups in the population under study. A false positive association can arise because allele frequencies and

disease prevalence vary across human populations (Figure 1.11).

The first attempt to solve this problem was the haplotype relative risk (HRR) approach [141]: the comparison of frequencies of marker alleles among cases and pseudocontrols (created from non-transmitted alleles). It is argued that the HRR method reduces, but not eliminate the possibility of population stratification [142]. Several other techniques have been proposed to deal with this issue: genomic control [143], structured association [144] and the use of principal components [145]. However, they all suffer from the same major disadvantage: they require a number (preferably >100) of widely spaced null SNPs that have been genotyped in cases and controls in addition to the candidate SNPs.

Figure 1.11 Example of True and False Association in Case/Control Study
(Taken from Hirschhorn et al. [146])

Shapes in red are the carries of a putative causative allele. In both figures (A. and B.) the fraction of individuals with the causative allele in cases is twice of that in controls. The upper part (A.) of the figure represents a true association: the frequency of risk allele is greater in cases than in controls in both populations. The bottom part (B.) of the figure represents a false positive association: the frequency of risk allele is identical in cases and controls in both populations, but because ethnic group 1 is overrepresented in cases and the risk allele is prevalent in ethnic group 1 only, the allele tested is overall twice as frequent in cases than in controls, leading to spurious association.



Unlike case/control studies, family-based analyses are robust against population substructure. Being "immune" to population stratification, the limitation of family-based TDT is the pedigree structure it can manage. Parents with an affected offspring (trios) are the recommended family structure for this test, since its chi-square statistic assumes that all observations are independent and this may not hold for pedigrees with multiple affected relatives [142]. It also requires knowledge of the genotypes of both parents, which in practice is not always feasible. Thus, TDT's original method has been expanded to suit diverse types of pedigrees. The new alternatives integrate the information carried by unaffected offspring to overcome the issue of missing parents by estimating their genotypes or by comparing the transmissions to affected offsrings with that to unaffected ones [147, 148]. The primary null hypothesis of no association and no linkage has been changed to "no association in the presence of linkage" to account for the non-independence of transmissions in cases of multiple affected relatives [147, 149]. Finally, splitting large pedigrees into nuclear families can be introduced to allow the analysis of pedigrees with multiple generations [147].

Extensions to TDT not only adjust the test to different sizes and types of pedigrees, but also make it possible to integrate a case/control study with a family-based analysis [147]. Such synthesis takes advantage of the strength of both approaches and represents a flexible alternative for association studies of complex traits.

1.2.4.3 Design and Interpretation of Genetic Association Studies

1.2.4.3.1 The Power of an Association Study

The probability of a study to obtain a significant result (power) is a critical aspect in the design of any genetic association analysis. It is, therefore, important to understand and evaluate the parameters influencing this statistical power, which include (1) the frequency and degree of risk attributed to a disease allele; (2) the sample size; (3) the degree of LD and allele frequency matching between a marker and a disease allele; (4) the mode of inheritance; (5) the prevalence of the disease and (6) the type of an association study. There is a complex interplay between all of these factors and one cannot be considered in isolation: the sample size required for a study, for example, greatly depends on the disease allele frequency and its relative risk [122, 150].

Under a family-based or case-control design of an association study, it is equally important to evaluate the disease allele parameters (direct study) and its degree of LD with a marker allele (indirect study) as they are the major determinants of the power through the sample size of a study: the greater the LD between a marker and disease allele, the greater the similarity between their population frequencies, and the greater the genotypic relative risk, the greater will be the power of an association test [150].

In indirect studies, the dominant feature of the relationship between a disease allele and a marker allele, however, is not LD but marker allele frequencies. Under the rare allele hypothesis, even if LD is complete, at least 1/3 of the true effect is lost with markers of 10% minor allele frequency or greater. The detection of a frequent disease allele suffers similarly from marker-related decay in effect size, but in the opposite direction: a marker with a minor allele frequency of 20% or less leads to very low effect sizes that can not be detected even with a vast sample size [122]. Thus, the minimum sample size is achieved when the frequencies of the disease allele and associated marker allele are equal [150].

The importance of equality between disease and marker allele frequencies becomes more apparent when the effect size of a disease allele is taken into consideration. In the situation of a disease with a common allele of small relative risk, a "common" marker allele with MAF as high as 50% will still result in a pronounced drop in effect size and, thus, power of an association study. The situation of a rare disease allele that also has small effect size will suffer from an even greater loss of power as the combination of the low effect size and rare MAF would lead to unfeasibly large sample size requirement to allow the detection of association[122].

1.2.4.3.2 Marker Selection for an Association Study

Performing any association analysis requires a selection of polymorphisms. However, this selection is not a simple matter. Firstly, for the powerful detection of a target polymorphism, all the variants in the population of interest should be represented, but in practice with current 300,000 to 1,000,000 SNPs genotyping platforms ~20% of common SNPs are only partially tagged or not tagged at all, and rare variants are generally missed out [116]. Secondly, the current maximum genotyping panel is one

million SNPs, while the estimated number of SNPs in the human genome is 15 million [89]. Thus, in actuality, markers are chosen based on LD patterns.

Based on the r^2 measure of LD, an $r^2 > 80\%$ is generally considered to be sufficient for tag SNP mapping to obtain a good coverage of untyped SNPs and relatively small loss in power [151]. The optimal number of variants for an association study is defined as the smallest number of SNPs that needs to be genotyped to cover the other SNPs at an r^2 of 0.8 or greater [138, 152]. If the LD between SNPs is strong, this could result in up to 70-80% less genotyping. In contrast, if LD in the region of interest is low, almost every SNP may need to be genotyped.

Testing the hypothesis of a gene being a disease-causing candidate, the SNP selection can also be led by the within-gene position of a polymorphism (a variant, for example, can be chosen from an intronic region that is conserved, and therefore, may present a functionally important regulatory sequence [153]) or by the virtue of its function (polymorphisms that alter function through nonsynonymous protein-coding changes, or through effects of translation [154]).

The above approaches are primarily aimed at surveying common variants (MAF > 5%). Rare polymorphisms require a more comprehensive analyses: considerably larger sample size and resequencing [155].

1.2.4.3.3 Interpretation of Results from an Association Study

An additional challenge of association studies is their interpretation. Most of the reported significant associations appear to be poorly reproducible (section 1.4). The possible explanations of this inconsistency are false-positive reports, misinterpretation or true heterogeneity between studies.

It is crucial to understand the nature of the association test performed (hypothesis generating or hypothesis testing), assess its power and correct for multiple testing (where applicable) before drawing conclusions.

In a hypothesis-generating genome-wide test, a number of polymorphisms are tested and the results can be considered significant only after appropriate correction for these multiple tests. It also should not be forgotten that these analyses are exploratory and their replication in an independent sample is necessary to distinguish false positives from true associations.

In candidate gene analysis (replication or hypothesis testing), it is important to choose the appropriate test according to the available sample of cases/controls or pedigrees. The wrong null hypothesis or method of family-based studies as well as the neglection of population stratification in case/control approach may all lead to false positives.

1.3 Myopia and Genetics

The link between myopia and genetics has been long recognized. Firstly, several familial studies report that myopia occurs more often in the children of myopic parents than non-myopic parents [156-159]. Yap et al noted that the prevalence of myopia in 7-year-old children increases up to 45% when both parents are myopic compared to 7.3% when neither parent is myopic [159].

Further, strong evidence for the role of genetics in myopia is also provided by twin studies: identical (monozygotic) twins display higher similarity in their refractive status and ocular components than dizygotic twins [30, 31, 34, 160].

Multiple familial studies (linkage and association) support the importance of a genetic effect on myopia. To date, 14 MYP regions linked to myopia are listed in the Online Mendelian Inheritance in Man (OMIM) and several novel intervals have been identified (Table 1.9). Determination of these loci has generally been based on just a few families with little replication of linkage by other investigators, except for the MYP3 region on chromosome 12, which has been replicated in 3 independent studies including one conducted by the International High Myopia Consortium with the largest dataset yet assembled [161]. Familial occurrence of myopia has been mostly described as a discrete, segregating trait based on the distinction of low and high grades of myopia, showing autosomal dominant inheritance in the majority of studies.

Although segregation analyses suggest the involvement of multiple genes, high myopia is more likely to result from a major effect mutation than are lower grades of refractive error [162].

The genetic intervals identified by linkage analysis harbour a number of loci encoding possible myopia genes. Several association studies – in which linkage peaks have been fine-mapped as well as candidate-gene analysis – have been carried out in an attempt to identify genetic variants that confer susceptibility to myopia (Table 1.10). Unfortunately, many of the initial reports of association proved to be false positives [163-169], leaving the majority of myopia susceptibility genes still to be discovered.

Table 1.9 Overview of Myopia Linkage Analysis Studies
(Abbreviations: AD: autosomal dominant; XR: X-linked, recessive; QTL: quantitative trait locus)

Locus	Original Finding					Replication				
	Location	Ethnicity of the Cohort	Myopia Criteria	Mode of Inheritance	Reference	Ethnicity	Myopia Criteria	Mode of Inheritance	Confirmation	Reference
MYP1	Xq28	Caucasian		XR	[170, 171]					
MYP2	18p11.31	Caucasian	< -6 D	AD	[172]	Caucasian	< - 6 D	AD	No	[173]
						Caucasian	< - 1 D	QTL	No	[174]
						Asian (Chinese)	< -6 D	AD	Yes	[175]
						Caucasian	< - 6 D	AD	Yes	[176]
						Caucasian	< - 5 D	AD	No	[177]
						Caucasian	< - 1 D	AD	No	[178]
MYP3	12q21-q23	Caucasian	< - 6 D	AD	[160]	Caucasian	< - 6 D	AD	Yes	[173]
						Caucasian	< - 1 D	QTL	No	[174]
						Caucasian	< - 5 D	AD	Yes	[179]
						Caucasian	< - 5 D	AD	Yes	[177]
						Caucasian	< - 1 D	AD	No	[178]
MYP4	7q36	Caucasian and African-American	< - 6 D	AD	[180]					
MYP5	17q21-q22	Caucasian	< - 6 D	AD	[181]	Caucasian	< - 6 D	AD	No	[173]
MYP6	22q12	Caucasian	< - 1 D	AD	[178]	Caucasian	< -1 D	AD	Yes	[182]
						Caucasian	Continuous	QTL	Yes	[183]
						Caucasian	< - 1 D	AD	No	[184]
MYP7	11p13	Caucasian	< -1 D	QTL	[185]	Caucasian	< -1 D	QTL	No	[186]
MYP8	3q26	Caucasian	< -1 D	QTL	[185]	Caucasian	< -1 D	QTL	Yes	[186]
MYP9	4q12	Caucasian	< -1 D	QTL	[185]	Caucasian	< -1 D	QTL	No	[186]
MYP10	8p23	Caucasian	< -1 D	QTL	[185]	Caucasian	< -1 D	QTL	No	[186]
						Caucasian	< - 1 D	AD	Yes	[184]
MYP11	4q22-q27	Asian (Chinese)	< - 6 D	AD	[187]					
MYP12	2q37.1	Caucasian	< - 6 D	AD	[188]	Caucasian	< - 0.5 D	AD	Yes	[189]
MYP13	Xq23-q25	Asian (Chinese)		XR	[190]	Asian (Chinese)		XR	Yes	[191]
MYP14	1p36	Caucasian	< - 1 D	QTL	[192]					
Novel	1q	Caucasian	Continuous	QTL	[183]					
Novel	7p21	Caucasian	Continuous	QTL	[183]					
Novel	5p15	Asian (Chinese)	< - 6 D	AD	[193]					
Novel	7p15	African-American	< -1 D	QTL	[194]					

Table 1.10 Overview of Myopia Candidate Gene Analysis
(Abbreviations: TF: transcription factors)

Gene Symbol	Gene Name	Reason for the Study	Cohort	Ethnicity	Myopia Criteria	Analysis Type	Significant Finding	Reference
TEX28	TestisExpressed28	Location within MYP1	5 Families	Caucasian	< - 5 D	Screening of sequence	Suggestive	[195]
NYX	Nyctalopin	Myopia in congenital stationary night blindness	52 Cases	Asian (Chinese)	< - 6 D	Screening of sequence	Suggestive	[196]
TGIF	Transforming Growth β-Induced Factor	Location in MYP2 and role in eye growth	71 Cases / 106 Controls	Asian (Chinese)	< - 6 D	Association of screened mutations	Yes	[197]
			204 Cases / 112 Controls	Asian (Chinese)	High	SNP analysis of exons	No	[164]
			10 Cases / 10 Controls	Caucasian	< - 6 D	Screening of sequence	No	[165]
			330 Cases / 330 Controls	Asian (Japanese)	<-9.25D	Association	No	[167]
			288 cases / 208 controls	Asian (Chinese)	< - 6 D	Association Study	No	[198]
			257 Cases / 294 Controls	Caucasian	< - 0.5 D	Association	No	[199]
			10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[200]
			10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[201]
CLUL1	Clusterin-like1	Location in MYP2	10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[200]
EMILIN2	ElastinMicrofiblrillnterfacer 2							
ZFP161	Zinc Finger Protein 161							
MYOM1	Myomesin 1		10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[201]
MRCL2 /3	MyosinRegulatoryLigt Chain 2 /3							
DLGAP1	DrosophilaHomolog Associated Protein 1							
LPIN2	Lipin 2	Location in MYP2 with highest LOD of 9.59[172]	10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[200]
			7 cases / 6 controls	Caucasian	< - 6 D	Examination of genomic structure, expression and SNPs	Potential regulatory elements for TF	[202]
DCN	Decorin	Location in MYP3 and role in collagen structure	10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of sequenced polymorphisms	No	[201]
			120 Cases / 137 Controls	Asian(Taiwanese)	< -10 D	Association test	No	[203]
EPYC	Epiphycan	Location in MYP3 and role in collagen structure	10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of screened polymorphisms	No	[201]

Table 1.10 Overview of Myopia Candidate Gene Analysis (Continuation)
(Abbreviations: TF: transcription factors)

Gene Symbol	Gene Name	Reason for the Study	Cohort	Ethnicity	Myopia Criteria	Analysis Type	Significant Finding	Reference
LUM	Lumican	Location in MYP3 and role in collagen structure	10 Cases / 10 Controls	Caucasian	< - 6 D	Segregation of screened polymorphisms	No	[201]
			10 Cases / 5 Controls	Caucasian	< - 6 D	Polymorphism analysis of screened sequences	No	[166]
			125 Cases / 308 Controls; 4 Families	Caucasian	< - 6 D	Association and segregation tests of screened variations	Yes	[204]
			120 Cases / 137 Controls	Asian(Taiwanese)	< -10 D	Association Study	Yes	[203]
			288 cases / 208 controls	Asian (Chinese)	< - 6 D	Association Study	No	[198]
FMOD	Fibromodulin	Role in collagen structure	125 Cases / 308 Controls; 4 Families	Caucasian	< - 6 D	Association and segregation tests of screened variants	Suggestive	[204]
			10 Cases / 5 Controls	Caucasian	< - 6 D	Screening of sequence	No	[166]
OPTC	Opticin	Role in collagen structure	125 Cases / 308 Controls; 4 Families	Caucasian	< - 6 D	Association and segregation tests of screened variations	Yes	[204]
COL1A1	Collagen, Type 1, Alpha 1	Location within MYP5 and relation to collagen	471 Cases / 623 Controls	Asian(Taiwanese)	< - 6 D	Association study	No	[169]
			330 Cases / 330 Controls	Asian (Japanese)	<-9.25 D	Association Study	Yes	[205]
			141 Families	Caucasian	< - 5 D	Association Study	No	[177]
COL2A1	Collagen, Type 2, Alpha 1	Relation to collagen	123 Families	Caucasian	<-0.75 D	Association Study	Significant	[163]
PAX6	Paired Box 6	Location within MYP7 and role in eye development	221 Dizygotic Twin Pairs	Caucasian	< -1 D	Linkage and Association	Linkage Only	[185]
			123 Families	Caucasian	<-0.75 D	Association Study	No	[163]
			164 Families	Asian (Chinese)	< - 6 D	Association Study	Yes	[206]
			188 Cases / 85 Controls	Asian(Chinese)	< - 6 D	Association Study	No	[207]
			596 Subjects	Caucasian		Association Study	No	[168]
			4 Pedigrees	Caucasian	< -5 D	Association Study	Suggestive	[208]
SOX2	SexDeterminingRegionYBox 2	Location in MYP8 and role in eye development	596 Subjects	Caucasian		Association Study	No	[168]
SOX2OT	SOX2 overlapping transcript	Location in MYP8 and role in eye development	1430 cases/ controls	Caucasian	< -1 D	Association Study	Yes	[186]
TGFβ1	Transcription growth factor beta 1	Possible role in axial elongation	330 cases / 330 controls	Asian (Japanese)	<-9.25 D	Association Study	No	[209]
			201 cases / 86 controls	Asian (Chinese)	< - 6 D	Association Study	Yes	[210]
			288 cases / 208 controls	Asian (Chinese)	< - 6 D	Association Study	No	[198]
HGF	Hepatocyte Growth Factor	Possible role in axial elongation	128 families	Asian (Chinese)	< - 10 D	Association Study	Yes	[211]
			288 cases / 208 controls	Asian (Chinese)	< - 6 D	Association Study	No	[198]
MMP3/ TIMP1	Matrix Metallopeptidase 3 /TIMP Metallopeptidase Inhibitor 1	Possible role in axial elongation	366 cases / 736 controls	Asian(Taiwanese)	< - 6 D	Association Study	No	[212]
MYOC	Myocilin	Location in linkage region 1q and possible role in myopic alterations	70 cases / 69 controls	Asian (Chinese)	< - 6 D	Association Study	No	[213]
			162 families	Asian (Chinese)	< - 6 D	Association Study	Yes	[214]
			97 cases / 92 controls	Asian (Chinese)	< - 8 D	Association Study	Yes	[215]
EGR1	Early Growth Response 1	Involvement in ocular growth	96 cases	Asian (Chinese)	< - 6 D	Mutation Screening	No	[216]

CHAPTER II.

MATERIALS AND METHODS

2.1 Recruitment and Sample Collection

Subject recruitment was carried out with ethical approval granted by Cardiff University Human Sciences Research Ethics committee (Cardiff, Wales, UK) and followed the principles of Declaration of Helsinki. Signed, informed consent was obtained from each participant.

The project aimed to recruit (1) families where high myopia is present and (2) unrelated individuals with or without high myopia (cases/controls). In order to recruit subjects, information about The Family Study of Myopia was placed online and sent out to optometrists/ophthalmologists. Patients of the Eye Clinic operating at Cardiff University were also approached.

Potential participants were sent an information pack (Appendix 1), containing detailed information about the project, a questionnaire and a consent form. Once subjects agreed to take part in the study, their subjective refraction was obtained from their optometrists/ophthalmologists and DNA samples were collected in the form of saline mouthwashes via post.

Each potential subject was routinely asked to perform two mouthwashes first thing in the morning (before eating, drinking or brushing teeth) in order to obtain maximum DNA yield [217]. The participants were requested to perform the mouthwash rinses twice, immediately one after the other, and to then post the mouthwashes back to our laboratory as soon as possible. Further details are given below.

2.2 Mouthwash as a Source of Human DNA

2.2.1 DNA Extraction from Mouthwashes

Participants were mailed 50ml skirted tubes containing 15-20 ml of sterile 0.9 % NaCl, and were asked to swish this vigorously in the mouth for 20-30 seconds, before spitting it back into the same tube.

On arrival back to laboratory, mouthwash samples were refrigerated for at least 40 minutes, and then centrifuged at 3500 rpm for 5 minutes in a Boeco C-28 centrifuge (Boeckel & Co, Hamburg, Germany). The supernatant was discarded, and the buccal cell pellet resuspended in 480 µl of Extraction Buffer (10 mM tris-HCl, pH 8.0, 1mM EDTA, 0.5% SDS) and frozen at -20°C until processed further. Upon thawing, 20µl of proteinase-K (10 mg/ml) was added to each cell suspension and incubated in a waterbath with continuous shaking (~100 rpm) at 37°C for 2 hours. To separate insoluble material, tubes were centrifuged at 14000 rpm for 3 minutes and the supernatant was transferred to a fresh Eppendorf tube containing ~25µl high vacuum grease (Dow Corning Ltd). The vacuum grease served as a barrier between the aqueous (DNA-containing) and organic (protein-containing) phases after phenol/chloroform (phenol:chloroform:isoamyl alcohol – 25:24:1) extraction was performed. Phenol-chloroform extraction was repeated up to twice more until the supernatant was clear. After the addition of 17µl 5M NaCl and 1ml 100 % ethanol, the DNA was precipitated overnight at -20°C and then centrifuged at 14000 rpm for 10 minutes. The supernatant was removed and the DNA pellet was washed with 1 ml ice-cold 70% ethanol. After air-drying for 3 minutes, the DNA pellet was dissolved in 50µl of TE (10 mM tris, 1 mM EDTA, pH 8.0) and incubated at 37°C for 30 minutes with periodic gentle mixing (full protocol is in Appendix 2).

To *quantify* DNA concentrations, spectrophotometry, fluorometry, UV transillminator gel imaging system and Polymerase Chain Reaction (PCR) can be used. To test the *quality* of DNA, gel imaging system and PCR may be applied. These techniques are described below.

2.2.2 Assessment of Mouthwash-extracted DNA

2.2.2.1 Spectrophotometry

Spectrophotometry measures the amount of light that a sample absorbs. A spectrophotometer operates by passing a beam of light through the compound in question and measuring the intensity of light reaching a detector. Different molecules absorb energy (light) at different wavelengths. For DNA, Ultra Violet (UV) light is applied. This UV wavelength can be absorbed by a number of molecules present in a sample. Hence, the absorbance method does not distinguish nucleotides, single stranded DNA or contaminants (e.g. proteins and trace amounts of phenol) from good quality double stranded DNA. Moreover, it is relatively insensitive and is not well suited for testing small volumes or concentrations of DNA [218-221]. Typical sensitivity is 150 ng/ml of double stranded DNA [222].

To quantify DNA, absorbance is usually measured at three different wavelengths: 260(A260), 280(A280) and 320(A320) nm. Light of 260 nm is the one absorbed most strongly by DNA and its value is important in the calculation of the concentration of DNA in a sample. The absorbance of a DNA solution with a concentration of 50 μg/ml at 260 nm is one unit, so the concentration of an unknown sample can be calculated by multiplying measured units of its absorbance by 50 and a dilution factor [222].

The A280 is used in a ratio of A260:A280 to determine the purity of DNA. Ratios below 1.8 signal the presence of contaminating chemicals (e.g. proteins, phenol). Absorbance at 320 nm (A320) provides information about proteins in a sample, since proteins absorb light of this wavelength, but DNA does not.

Spectrophotometry is probably the most widely applied method for DNA quantification, but is limited by requirement of large sample volumes (~100 μl), poor detectability and lack of DNA specificity. Free nucleotides, single-stranded nucleic acids (e.g. ribonucleic nucleic acid) and proteins may exhibit significant absorbance at A260 and any contamination of sample preparation by these agents will result in over-estimation of the DNA concentration [223]. In addition, it has been shown that such factors as pH or

presence of phenol in the sample solution have a significant effect on the A260/A280 ratio [224, 225]. Therefore, only highly purified DNA preparations can be accurately quantified by spectrophotometry.

2.2.2.2 Fluorometry

2.2.2.2.1 The Principle of Fluorescence and Florometry

Fluorometry is the measurement of fluorescence, which is the phenomenon of light emission by "excited" molecules. Fluorescent molecules (fluorophores) absorb light at one wavelength and emit light at another. When fluorophores absorb light of a specific wavelength, their electrons rise to a higher energy level (the excited state). Electrons in this state are unstable and return to the ground level, releasing energy in the form of light. This emission of energy is fluorescence [226].

Analytical tools based on fluorescence are very useful because of their sensitivity and selectivity. When an analyte is fluorescent, direct fluorimetric detection is possible by means of a spectrofluorimeter operating at appropriate excitation and observation wavelengths. However, most molecules, including DNA, are not fluorescent and an indirect method of a fluorescent complex formation is applied in their analyses [226].

2.2.2.2.2 Fluorophores

Fluoresce-based analyses of nucleic acids are an integral part of many molecular biology procedures: fluorometry, agarose gel electrophoresis and real-time PCR.

The most commonly used fluorophore is ethidium bromide (EtBr), which is reported to have a sensitivity limit of 1ng/band for double stranded DNA (dsDNA) in agarose gels [227]. However, EtBr is potentially carcinogenic, posing handling and disposal problems. Furthermore, it easily photobleached and has a low fluorescence enhancement upon DNA binding, leading to high background readings [228, 229]. To address these issues, a series of cyanine dyes – such as SYBR green [230], PicoGreen [218, 229] and SYBR gold [231]

- have been developed. As a group, these dyes are characterized by having specific and high binding affinity to nucleic acids (up to 4x more than EtBr [231]), a low intrinsic fluorescence and large fluorescence enhancements upon binding to dsDNA [218, 228-231]. When bound to dsDNA, little background occurs since the unbound dye has virtually no fluorescence.

There are, however, certain limitations in their use as well. Cyanine dyes, for example, are not human-specific as they cannot distinguish between DNA molecules of different nature.

2.2.2.3 Ultraviolet (UV) Transillminator Gel Imaging System

2.2.2.3.1 Concept of DNA Gel Electrophoresis

Gel electrophoresis is a method that separates macromolecules on the basis of size, electric charge and other physical properties. The process of electrophoresis refers to the electrical charges "carried" by the molecules [232].

Nucleic acids are negatively charged. Under the influence of an electric field they migrate towards the positive electrode. The medium (e.g. agarose gel) they move through and their overall shape both affect their progress. It follows that different sizes and forms of nucleic acids move at different rates, providing the basis for their separation [232].

The basic protocol for DNA agarose gel electrophoresis can be divided into three steps: (1) a gel is prepared with agarose concentration appropriate for the size of DNA fragments; (2) the DNA samples are loaded into the wells of a gel and are run at a voltage and for time period that will achieve optimal separation; (3) the gel is stained or, if the dye was incorporated into the DNA sample, visualized directly upon illumination with UV light [233].

Agarose gel electrophoresis can be used as a quantifying and/or quality assessment method. To calculate the concentration of nucleic acid in an agarose gel, an image analysis

computer program is applied. DNA quality can be evaluated by examining the size and shape of bands. The next two sections describe the methods of DNA staining, visualization and evaluation in electrophoresed agarose gel.

2.2.2.3.2 Staining methods for DNA Gel Electrophoresis

In order to detect DNA bands in electrophoresed gel, such methods are applied as fluorescence and staining with silver or visible organic dyes.

Fluoresce-based visualization of nucleic acids is an integral part of digital fluorescent imaging is widely used for both documentation and analysis of electrophoretic separations of DNA. Fluorophores that aid such examination of DNA are described in section 2.2.2.2.2. Ethidium bromide and cyanine dyes are most widely used fluorescence reporters. It is possible to load the dyes directly to the DNA sample or agarose gel, avoiding the step of staining after a gel has been electrophoresed. However, all of the different types of dye can alter electrophoretic mobility and, thus, DNA size estimates [231, 234].

Silver staining [235, 236] requires a large number of laborious processing steps involving accurate timing [237]. Although it has been reported to be more sensitive than EtBr [235], silver staining is still less sensitive than SYBR green and also is expensive [238].

Organic visible dyes are simple and safe to use, but a long destaining step (more than one hour) to detect distinct DNA bands and low sensitivity (2- to 4-fold less than EtBr) limits their application in molecular biology [239]. Nonetheless, these dyes could be a plausible alternatives as their inclusion in agarose gels allows observation of DNA bands in ambient light, eliminating the application of damaging UV light required by fluorophores.

2.2.2.3.3 DNA Quantification with UV TRansilluminator System

Quantification of the amount of DNA in a sample is a critical step in wide selection of molecular biology experiments. One of the ways of measuring DNA quantity in a sample

is the densitometric analysis of bands with unknown DNA concentration run on a gel alongside standards of known concentration. The DNA can be quantified by constructing a "standard curve" and use of linear regression (section 2.4.1.2).

Fluorescence of the DNA-dye complex can be detected with an ultraviolet transilluminator system. This instrument represents a "dark room" where the gel is exposed to high intensity UV light and the induced fluorescence is captured by an attached digital camera. The image of fluorescent bands can be recorded on a disc and analyzed later using image analysis software.

Digital fluorescent imaging has a number of advantages, including (1) DNA specificity, because of the dye used to visualize DNA; (2) the ability to show the quality of DNA (if it is degraded, there will be fragments of different sizes after separation) and (3) application to a wide range of stains, as most of DNA fluorophores have excitation peaks with UV light. However, this method also has some drawbacks: variation in such factors as gel thickness, sample loading volume and DNA fragment size can have a relatively large effect on the fluorescent signals seen with equivalent amounts of DNA [240].

2.2.2.4 Polymerase Chain Reaction (PCR)

2.2.2.4.1 Conventional PCR

DNA molecules can be "mass-produced" from incredibly small amounts with the Polymerase Chain Reaction (PCR) technique. This discovery allows a researcher to mimic the cell's own natural DNA replication process in a test tube.

The PCR method uses specially designed DNA oligonucleotides (primers) that are complementary to the part of sample DNA to be amplified. The sample DNA is denatured by heating and upon cooling this allows the primers to bind to their target sequences, if present. In the presence of a suitably heat-stable DNA polymerase and DNA precursors (the four deoxynucleoside triphosphates - dNTPs), the bound primers initiate the synthesis

of new DNA strands which are complementary to the individual DNA strands of the target segment [241].

The PCR is a chain reaction because newly synthesized DNA strands act as templates for further DNA synthesis in subsequent cycles. After about 25 cycles of DNA synthesis, the products of PCR will include enough (about 10^5) copies of the specific target sequence to be easily visualized as a band of particular size when submitted to agarose gel electrophoresis [242].

PCR can provide information about DNA quantity (section 2.2.2.4.3) and quality. To get successful amplification one needs to have a DNA of a good quality: nicked or degraded DNA will not serve as a template for PCR.

2.2.2.4.2 Real Time PCR

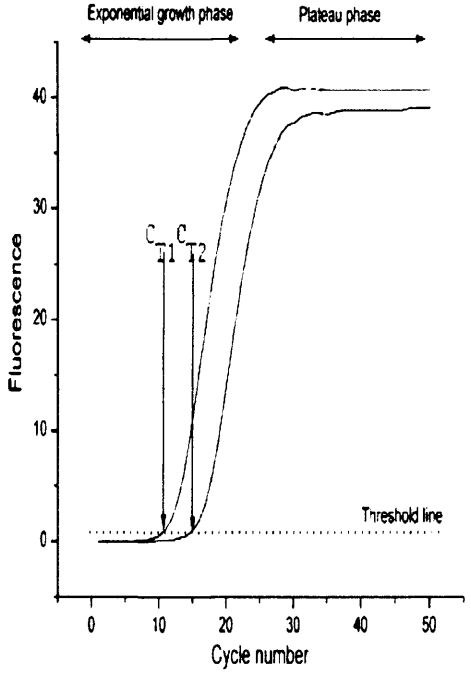
Conventional polymerase chain reaction has several limitations. In terms of DNA quantification, the biggest issue is that the reaction reaches a "plateau" phase, after which the PCR yield remains constant [243]. The ideal solution to this problem is a real-time PCR – a method that allows the detection of DNA sequences simultaneously with their amplification, first developed by Higuchi et al [244].

During the course of a real-time PCR reaction, detection of PCR products is made possible by including in the reaction a fluorophore (section 2.2.2.2.2) that reports the amount of DNA (this will yield a proportional increase in the fluorescent signal with the number of cycles). The information obtained is an amplification curve (Figure 2.2). This curve reflects two main phases of fluorescence: (1) an exponential growth phase when the product approximately doubles providing the efficiency of the reaction is 100%, and (2) a plateau phase when the reaction saturates and no increase in fluorescence can be detected (Figure 2.1). In a typical real-time PCR all curves saturate at the same level and, thus, the end point of a reaction can provide no information about the initial DNA concentration. The growth phase, on the other hand, provides information regarding the original concentration of template. The number of cycles needed to accumulate enough product to

raise the fluoresce signal above the background level is called the threshold cycle (C_T). Measurement of C_T is the quantitative basis of real-time PCR [243].

Apart from being able to monitor amplification, real-time PCR can also provide information about the PCR product itself, by means of melting it and registering the decrease in florescence as the dye is released from denaturing dsDNA. The temperature dependence of the fluorescence reduction is represented as a melting curve (Figure 2.2). The melting temperature of a product (T_m) is defined as the point at which 50% of DNA is double stranded and 50% is single stranded, and is a function of product size and base composition [245]. T_m can also be identified as a peak value in the negative derivative melting curve (Figure 2.2). The melting curve analysis can be used for quantification (as the area under the curve of the peak is proportional to the amount of product [246]) or for confirmation of the correct target sequence (non-specific products have different length and therefore deviating melting temperatures [245]).

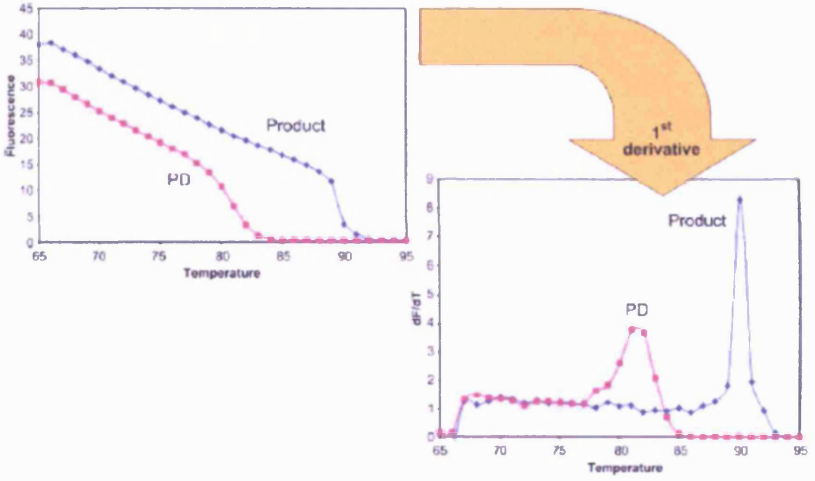
Figure 2.1 Real-time PCR Amplification Curve
(Taken from Kubista et al [243])



During the exponential growth phase the amount of PCR product approximately doubles in each cycle. However, as the reaction proceeds, the amplification slows and enters the plateau phase.
Initially, the fluorescence remains at the background level and its increase is not detectable. The time needed to accumulate enough product to yield a detectable signal is dependent on the initial DNA concentration and, thus, is different for each sample (C_{T1} and C_{T2}), providing the basis for the quantitative aspect of real-time PCR.

Figure 2.2 Melting Curve of a Real-Time PCR
(Taken from Kubista et al [243])

The figure shows the drop in fluorescence as the product/primer-dimer (PD) melts. The T_m is determined as the inflection point of the curve, which is easier to identify as the maximum peak in the negative first derivative of the melting curve.



2.2.2.4.3 Kinetics of Polymerase Chain Reaction and DNA Quantification

The basic equation describing PCR in simple terms is $N_c = N_o \times (E+1)^c$, where "c" is the number of thermocycles, "E" is the efficiency of a reaction, N_c is the amount of new product and N_o is the initial number of template molecules. This equation reflects the kinetics of PCR: each cycle produces an increase in N_c in proportion to amplification efficiency [247]. Since extension products are complementary to and capable of binding primers, each successive cycle essentially doubles the amount of DNA synthesized in the previous cycle. This results in exponential accumulation of amplicon, approximately 2^n, where *n* is the number of cycles [248]. Therefore, 100% efficiency produces a doubling in specific target fragment.

The exponential growth in PCR product is not an unlimited process. Eventually, there will be more primer-template substrate accumulated than the amount of enzyme present is

capable of completely extending in the amount of time allowed. When this occurs, the efficiency of reaction declines and reaches a plateau phase when the product accumulates in a linear rather than exponential manner [248]. In addition, as the PCR reaction progresses and the initial molar excess of primers present starts to reduce, template-template re-annealing can out-compete the primer-binding, leading to the plateau of a reaction.

Rearrangement of the PCR equation to $N_o = N_c / (E+1)^c$ provides the mathematical relationship upon which the quantitative PCR is based. Thus, quantification of N_c allows the calculation of N_o if amplification efficiency is known.

As described above, calculation of the initial DNA quantity in a sample with real-time PCR can be achieved using the C_T method, in which individual reactions are compared at the point when they contain identical amounts of product: at the C_T threshold N_c becomes constant and, thus, $N_o = N_T / (E+1)^{Ct}$ [249].

The efficiency of a PCR assay (as well as the concentration of unknown samples) can be estimated from a standard curve based on serial dilutions of a standard sample. The C_T values of diluted standards are read out and plotted against the logarithm of their concentrations or dilution factor. The mathematical basis of a standard curve can be derived by taking the logarithm of the C_T method equation: $\log(N_o) = \log(N_T) - \log[(E+1)^{Ct}]$, which can be rewritten as $\log(N_o) = -\log(E+1) \times C_T + \log(N_T)$. Assuming that E and N_T are constant, this standard curve equation will be linear and, therefore, plotting $\log(N_o)$ versus C_T will produce a line with a slope of $-\log(E+1)$ and intercept of $\log(N_T)$. Hence, PCR efficiency can be calculated from the slope of a standard curve as $10^{-1/Slope}$ [247]. Amplification efficiency is also frequently presented as a percentage: the percent of template that was amplified in each cycle. To convert E into a percentage the following equation can be used: Efficiency(%) = (E-1)x100% [250].

For proper comparison and quantification of DNA, all samples should have similar amplification efficiency during the exponential phase of a reaction: even as small a difference as 5% will result in a 3-fold difference in the amount of DNA after 25 cycles of

exponential amplification [251]. In addition, C_T values generated from different runs can be compared directly only if an identical threshold was used for each run. Finally, the relationship between N_T and threshold values is dependent on amplicon size because the DNA fluorescence that underlines the determination of a threshold has a linear relationship with DNA mass [247].

The accuracy of DNA quantification in biological samples is often difficult because of their complex nature: they may contain inhibitory substances that are not present in purified standards. These inhibitory factors include carry-over chemicals from DNA purification, detergents, antibiotics, buffers, enzymes, fats and proteins [252]. The presence of an inhibitor may result in apparent increase in efficiency: samples with the highest concentration of template also have the highest amount of inhibitors, which causes a delayed C_T; whereas samples with lower template concentration have lower levels of inhibitors, so their C_T is minimally delayed. As a result, the absolute value of slope decreases and calculated efficiency appears to increase [250].

2.3 Genotyping

2.3.1 Microsatellite Genotyping

Precise and reproducible sizing of DNA fragments generated by PCR has become a fundamental technology in microsatellite genotyping. The procedure essentially involves two steps: amplification and electrophoresis [97]. DNA polymorphisms are amplified with end-labelled primers and visualized after separation on denatrating polyacrilamide sequencing gels [97]. The original polyacrilamide gel has been replaced by capillary electrophoresis in small diameter tubes [253] as it has benefits of automated filling of the capillary with separation medium and automated sample loading, allowing for full automation of the process.

One of the most widely used systems for microsatellite genotyping is the ABI 310 Genetic Analyzer. On this system, polymorphic loci are amplified with one unlabelled and one fluorescently 5'-labelled primer. Denaturated PCR products are then electrophoresed with

an internal size standard (DNA fragments of known size labelled with a different fluorophore). Multiplex products are sequentially injected into a single capillary and detected in real time as they pass by a laser-detection window during their electrophoresis. The laser-induced fluorescence is captured with a CCD camera. The collected data is then analyzed by software that manually or automatically determines allele sizes on the basis of a standard curve from the internal size standards [254].

Precise scoring of microsatellite alleles, regardless of repeat unit size (2, 3 or 4 nucleotides), holds a number of requirements. Genotyping errors may arise in several ways: low quality/quantity template DNA, unreliable PCR amplification or incorrect calling of alleles. Therefore, quality control is an essential step in accurate microsatellite genotyping (section 2.3.3).

2.3.2 SNP Genotyping

SNP genotyping is a major part of any large-scale genetic study and, thus, an appropriate genotyping method is crucial. The ideal assay should be sensitive, robust, automated and cost-effective. The majority of protocols involve the following steps: (1) allelic discrimination chemistry (hybridization, flap endonuclease discrimination, primer extension, allele specific digestion and oligonucleotide ligation); (2) allele detection (monitoring of the light emitted by products, measuring the mass of products or detecting a change in electrical property when the product is formed) and (3) allele calling. The challenge of high-throughput genotyping lies in pairing the right chemistry assay with the right detection system to maximize efficiency with respect to accuracy, speed and cost.

Hybridisation chemistries coupled with fluorescent plate reader detection currently offer the simplest route to a high-throughput genotyping platform. In this method allele-specific probes are immobilized on a solid support to capture amplified, labelled target DNA samples, and the hybridization event is visualized by detecting the label after the unbound targets are washed away. Knowing the location of the probe sequence on the solid support allows inference of the genotype of the target DNA. Hybridization assays differ in their way of reporting allele-specific binding: (1) TaqMan monitors the cleavage event of a

specific probe during PCR [255]; (2) with molecular beacons, detection is based on the fluorescence of a stem-loop structure upon binding to the target DNA [256]; (3) in the "light-up" technique, fluorescence of the target DNA - oligomer probe complex reports the hybridization [257].

Allele-specific hybridization is the basis for an elegant genotyping assay: a complete system to generate a large number of short PCR products for each SNP in multiplex amplification and to automate hybridization, data scanning and analysis, allowing the screening of a large number of SNPs in parallel (e.g. Illumina or Affymetrix genotyping platforms). The major advantage of performing genotyping reactions on solid supports (e.g. latex bead, glass slide or silicon chip) is that many markers can be interrogated at once, saving time and reagents. Common hybridization conditions used for multiplexing, however, pose a problem of not being specific for all of the genotyped SNPs, with subsequent implications for data quality (section 2.3.3).

2.3.3 Genotyping Errors and their Prevention/Detection

Pinpointing genetic associations relies heavily on the accuracy of the underlying genotype data. High-throughput genotyping errors may occur for a number of reasons: low quality/quantity of template DNA, unreliable PCR amplification, electrophoresis artefacts, assay non-specificity, incorrect calling of alleles and data entry errors.

Erroneous PCR amplification can be caused by deficient template DNA, poor primer design or suboptimal reaction conditions. Under any of these circumstances it is possible that one allele of a heterozygote will not be detected and that false allele calls will arise [258, 259]. Two particular problems are often experienced with microsatellite genotyping: (1) stuttering (minor products preceding the primary allele peak on electopherogram) and (2) an extra adenine base (A) added to the 3' end of the amplified product by Taq polymerase. Both of these artefacts can cause difficulties in allele calling, particularly when analyzing dinucleotide repeats.

Nonrandom genotyping failure, which involves an individual's SNP genotype that is either incorrectly called or more commonly not called at all, can be a source of confounding in genome-wide association analysis. If such failure is non-random with respect to genotype (e.g. some genotypes are more likely to be uncalled) and to phenotype (e.g. cases have lower genotyping rates than controls) then false positive association can occur [260]. Therefore, it may be beneficial to exclude those SNPs or subjects that show a high genotyping failure rate.

Quality control and accurate quantification of DNA samples, as well as reproducibility checking by running replicates of samples of known genotypes, may prevent faulty genotyping. To account for an extra A when dealing with microsatellites, a high single-base resolution genotyping method is required.

To detect genotyping errors, Mendel consistency tests and HWE checking can be performed. Alleles showing non-Mendelian behaviour in families or out of HWE among unrelated subjects should be re-called (e.g. in pedigrees) or excluded from further analyses.

2.4 Statistical Analyses

2.4.1 Generalized Linear Model

One of the most widely used tools of statistical analyses is the generalized linear normal model, exemplified by analysis of variance (ANOVA) and by regression analysis (sections 2.4.1.1 and 2.4.1.2).

Any statistical test of pattern requires a model against which to test the null hypothesis of no pattern. A linear model analyzes the relationship between two variables: one independent (or explanatory) and one dependent (or response). A model often comes in the form of a numerical function of input variables. Apart from the independent (input) variable, the function also contains some numerical parameters that need to be adjusted

to the data by some type of algorithm.

The generalized normal linear model requires four basic assumptions: (1) the dependent variable is normally distributed, (2) the variance of the dependent variable remains constant over the range of values of the independent variable to be considered, (3) the mean of the dependent variable is a linear function of any parameters introduced and (4) the observations of the independent variable are independent [261]. Thus, before applying any linear model to a data set, normality and homogeneity of variances must be addressed.

2.4.1.1 Analysis of Variance (ANOVA) and Post-Hoc Tests

ANOVA aims to identify whether there is any significant difference between the means of two or more groups of data. However, it does not compute the differences between means of the groups directly. Instead, ANOVA focuses on the variability in the data, examining if the variance between the group means is greater than would be expected by chance [262]. Thus, ANOVA is termed analysis of variance.

The statistic of ANOVA is the F-ratio or F-statistic: the value is obtained from the ratio of the variance between the groups and the variance within the groups [263] . A variance is the measure of variability, taking account of the size of the dataset.

The variability in a set of data quantifies how different the individual observations are from the mean of the overall population in general. Therefore, before performing an ANOVA test, it is crucial to assure the homogeneity of variances in groups of data wished to be analyzed (e.g. Levene's test).

To calculate the variance, first the grand mean of the population is calculated, then the differences of each point from the mean: deviations will be both positive and negative, and the sum will be zero. This will hold regardless of the size or the amount of variability of the dataset. Thus, the raw differences are not useful as a measure of variability. If the measures are squared before summation, on the other hand, then this sum is a better

estimate of variability: it will increase the greater the scatter of the data point around the mean. This quantity is called the sum of squares (SS) and is the basis of the F-statistic [262].

The total sum of squares can be divided into two parts: SS between groups (SS_B) and SS within groups (SS_W). SS_B is calculated based on the squared deviations between the group means and the grand mean of the sample population (all groups together). SS_W is calculated based on the squared deviations between each individual in a group and that individual's group mean [262].

The SS, however, cannot be used as a comparative measure between groups because it will be influenced by the number of datapoints in the group: the more datapoints the greater the SS. Therefore, SS is converted to a variance or the mean square (MS) by dividing SS by degrees of freedom [262].

In statistics, degrees of freedom represent the number of independent pieces of information in a population of size n required to obtain a given grand mean. Since all deviations must sum up to zero by definition, it is known what the final value must be and, thus, there are only $n-1$ independent observations or degrees of freedom [262, 263].

The mean square between the groups (MS_B) is the sum of squares between groups (SS_B) divided by the degrees of freedom between groups (DF_B), and, similarly, the mean square within the groups is SS_W divided by degrees of freedom within the groups (DF_W). Degrees of freedom between the groups is one less than the number of groups, while DF_W is the difference between the total degrees of freedom ($n-1$) and DF_B. The F-statistic is then the ratio of the mean square between groups and the means square within groups ($F= MS_B / MS_W$) or it is between groups mean square divided by the error mean square [262, 263].

Being a type of general linear model analysis, ANOVA is analogous to the regression situation with the mean of the independent variable forming the fitted value and the following equation: total deviation = deviation explained by independent variable + unexplained deviation (residual) [263].

If the p-value of the F-ratio or ANOVA test is less than 0.05, it only indicates that there is significant discrepancy between groups and does not give any information on which specific means are different. To uncover the source of significance, so-called post-hoc tests need to be performed [262].

The appropriate post-hoc examination is dependent upon the number and type of comparisons planned. The only post-hoc analysis used in this study is Dunnett's test, which is used for comparability of groups with a chosen reference group, such that there are one less comparisons than the total number of groups [262].

2.4.1.2 Linear Regression

The purpose of linear regression analysis is to evaluate the impact of a predictor (independent) variable on an outcome (dependent) variable [264].

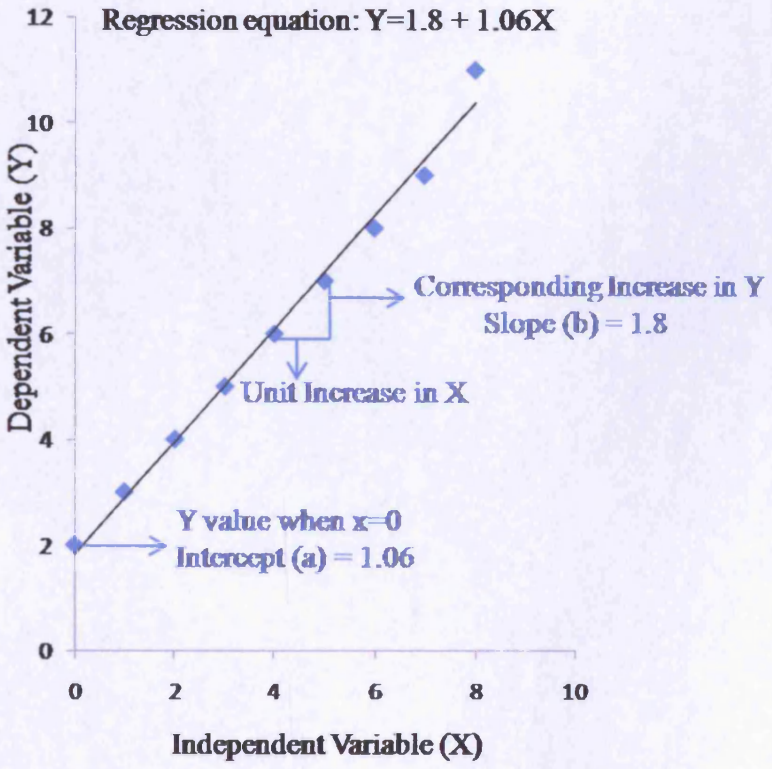
A simple, univariate regression model contains only one independent (explanatory) variable (x) and is linear with respect to the dependent variable. Mathematically, the model is expressed as $Y = a + bx + e$, where "Y" is the outcome variable; "x" is dependent variable; "a" and "b" are parameters of the model, representing the intercept on y axis and slope of the regression respectively; and "e" is the random error [264] . The slope is the average change in Y if x were to change by one unit and the intercept is the Y value when x equals zero [265] (Figure 2.3).

The main goal of linear regression is to fit a straight line through the datapoints, that would best explain the relationship between the two examined variables. This regression line is obtained using the method of least squares. Any line drawn through the datapoints would give a predicted or fitted value of Y for each observed value of x in the data set. The vertical difference between the observed and fitted value of y for a particular point x is known as the deviation or residual. The method of least squares finds the values of "a" and "b" that minimise the sum of the squares of all deviations [264, 266].

The fraction of the variability in Y (with a range of 0-1) that can be explained by variability in x through their linear regression is referred to as the R-square measure. [264].

Several hypotheses can be tested by linear regression. The most common one is to test whether the slope of the regression line is zero (there is no linear relationship between Y and x). However, it is important to understand that linear regression should not be interpreted as causation and should not be used to estimate outside the range of the independent variables [264].

Figure 2.3Graphical Representation of Linear Regression
(Taken from Zou et al [264])
Blue squares represent the datapoints, across which a line was fitted using linear regression. The expectation of the dependent variable Y is linear in the dependent variable x, with an intercept a = 1.06 and a slope b = 1.8.



2.4.2 Analysis of categorical outcomes

2.4.2.1 Statistics of Contingency Tables

2.4.2.1.1 Concept of Contingency Tables

When using categorical (qualitative) variables in an investigation of statistical relationship between groups, the data can be summarized in form of frequency or counts of independent observations in each category. If a statistical test is restricted to the association between two dichotomous variables, then the counts can be presented in a 2x2, or contingency, table [267, 268] (Table 2.1). It is important to assure that the outcome for each individual is independent of the outcome for other individuals in order not to violate the assumption of independent observations of the statistics of contingency tables.

Table 2.1 Contingency Table (Example)
(Taken from Sistrom and Garvan [268])

Both dependent (outcome) and independent (risk) variables are dichotomous. Counts of observations are calculated for each of the two categories of each variable. N is the sum of all observations.

		Outcome (Dependent) Variable		**Total**
		Category One	Category Two	
Risk (Independent) Variable	Category One	a	b	a + b
	Category Two	c	d	c + d
Total		a + c	b + d	N

2.4.2.1.2 Measures of the Effect of a Risk Variable in Contingency Tables

The *probability* of the occurrence of a particular event equals the proportion of times that the event would (or does) occur in an examined population. For example, the probability of death in five years following diagnosis of prostate cancer would be defined as the proportion of times death would occur among a large number of men diagnosed with prostate cancer. This probability is then said to be the *risk* of death in the five years following the diagnosis of prostate cancer [269].

The probability has a value between 0 and 1: 0 if the event never occurs and 1 if it is certain to occur. It can also be expressed as a percentage, taking a value between 0% and 100%.

The *odds* of an event A are defined as the probability that A does happen divided by the probability that it does not happen: Odds(A) = prob(A) / 1 –prob(A). The odds are always bigger than the probability since 1 – prob(A) is less than one: for example, when the probability is 0.5, the odds are one (0.5/(1-0.5)). In contrast to the probability, which lies between 0 and 1, odds take the rage from 0 (when prob(A) = 0) and infinity (when prob(A) = 1). When the probability is small (<0.1), the odds are very close to the probability because 1-prob(A) would be very close to one [269].

The effect of a certain risk variable can be assessed using probability and odds in three ways: risk difference, relative risk (RR) and odds ratio (OR) (Table 2.2).

The risk ratio is more commonly used to measure the strength of an association than is the difference in risks. This is because the amount by which a risk factor multiplies the risk of an event is interpretable regardless of the size of the risk. A risk ratio of one occurs when the risks are the same in the two groups and is equivalent to no association between the risk factor and the outcome. A risk ratio greater than one occurs when the risk of the outcome is higher among those exposed to the risk factor than among the non-exposed [269].

An odds ratio of 1 occurs when the odds, and hence the proportions, are the same in the two groups and refer to no association between a risk factor and an outcome variable. As the risk in the group with no risk becomes larger, the maximum possible value for of the risk ratio becomes constrained, because, by definition, it must not be more than one. Odds ratio, on the other hand, is not constrained in this matter since there is no upper limit to its value [269].

Both risk ratio and odds ratio reflect the ratio of proportions and, thus, the hypothesis of it being equal to one or not can be tested. If the calculated ratio's 95% confidence interval does not include one, it means that the OR or RR show a significant effect of the examined risk variable [270].

Table 2.2 Measures of the effect of a risk variable in a contingency table
(Adopted from Kirkwood and Sterne [269])

The notations "a", "b", "c" and "d" are the same as in Table 2.1.

Measure of comparison	**Formula**
Risk Difference	$a/b - c/d$
Risk Ratio (Relative Risk)	$\frac{a / b}{c / d}$
Odds Ratio (OR)	$\frac{a \times d}{b \times c}$

2.4.2.1.3 Chi-square and Fisher Tests in Contingency Tables

In order to test for association in two categorical variables organized in a contingency table, the χ^2 (chi-square) or Fisher's test can be performed. The χ^2 test makes comparison between the observed or collected data versus the data one expects to find: $\chi^2 = \Sigma(\text{observed-expected})^2/\text{expected}$. In other worlds, the test examines if the difference between observed and expected values is due to random chance or some factor influencing the results [267].

The use of the "chi-square distribution" in tests of association is an approximation that relies on large expected frequencies and, thus, the cell counts cannot be less than five [267, 270]. To overcome this limitation, the relationship can be tested by Fisher's exact test, which evaluates the probability of obtaining the particular, observed cell counts, considering the total number of all possible tables with the given marginal totals and assuming the null hypothesis of no association [267, 270].

For large sample sizes the two statistics give very similar results, but for smaller samples Fisher's test is preferable, although being more conservative (Fisher's test produces larger p-values with less probability to conclude significant association between studied variables [270]).

2.4.2.2 Logistic Regression

The basis of the simple logistic regression is derived from the odds ratio (OR) of an examined risk factor: Odds in exposed group = Odds in unexposed group x Odds ratio. This model expresses the odds in each group in terms of two model parameters: baseline and odds ratio. The term baseline refers to the group against all other groups will be compared, while odds ratio expresses the effect of a risk factor on an outcome variable. Because confidence intervals or odds ratios are derived by using the *log* function, logistic regression models are fitted on a logarithmic scale. Thus, the previous equation can be rewritten as log(Odds) = log(Baseline) + log(Odds Ratio), transforming it from multiplicative to an additive (or linear) one [269].

The general form of the logistic regression model is similar to that of the linear regression: $\log(\text{Odds}) = \beta_o + \beta_1 x$, where β_o and β_1 are the regression coefficients, and x is an independent (or exposure) variable. For comparing two exposure groups the exposure variable would equal one for those in the exposed group and zero for those in the unexposed group [269].

The likelihood ratio statistic in the logistic regression is the so-called Wald test, which is based on a quadratic approximation of the exact log likelihood ratio, chosen to have the same value and curvature at the maximum likelihood estimate [269] (section 2.4.3)

2.4.3 Likelihood and Likelihood Ratio

The likelihood gives a comparative measure of how compatible is an examined dataset with each particular value of a probability. For example, after testing 12 households for tuberculosis, 3 tested positive and 9 tested negative. Using the notation of Table 2.1, (a + c) would equal 3 and (b+d) would equal 9. The sample proportion, thus, would be 0.25. The likelihood would give the value of the most likely probability of trasmitting a tuberculosis infection given the sample proportion of 0.25 [269].

The approach used to calculate the probability (π), or *likelihood*, is the maximum likelihood estimation (MLE) and is derived by differentiating the binomial likelihood equation of $\pi^{(a+c)} \times (1-\pi)^{(b+d)}$ to find the value that maximizes it. The result is (a+c)/(a+c+b+d), which is (a+c)/N and in this example is 0.25 [269].

As well as concluding that 0.25 is the most likely value for the true probability π of the risk of household transmission of tuberculosis in this example, it is useful to know what other values of π are compatible with the data. The likelihood for any other probability will be less than MLE. How much less likely is assessed using the *likelihood ratio (LR)*: LR = Likelihood for π / Likelihood at MLE. By definition, the likelihood ratio equals one for the MLE and less than one for all other values [269].

Because the confidence interval for the likelihood ratios is derived using logarithmic function, in practice log(LR) is calculated instead of the actual ratio. Provided the sample size is sufficiently large, the curvature of log(LR) can be approximated by a quadratic equation, which is easier to handle mathematically.

Likelihood is usually used for hypothesis testing and there are three types of tests based on the log likelihood: the likelihood ratio test (LRT), Wald test and score test. The likelihood ratio test (LRT) is based on the value of the log likelihood ratio at the null value of the parameter and equals -2 x ln(Likelihood at null parameter – Likelihood at MLE) [269].

The Wald and score tests are both based on the value of a fitted quadratic approximation. The Wald test uses the approximation to the log likelihood ratio at the null value of the parameter of interest rather than the actual value of the log likelihood ratio at this point. The quadratic approximation of the Wald test is chosen to meet the log(LR) at the MLE point and to have the same curvature as the log(LR) at this point. It is symmetrical around MLE and its maximum value is zero [269].

The score test, on the other hand, uses an alterative approximation, chosen to have the same value, gradient and curvature as the log likelihood ratio at the null value of the parameter rather than at its MLE [269].

2.4.4 Statistical Methods for Detection of Genetic Association of Refractive Error

2.4.4.1 Refractive Error as a Categorical Outcome

Refractive error can be assessed both as categorical or continuous variable: continuous tests would provide information on refractive error as such; while performing association tests on dichotomized – affected (highly myopic) and non-affected (emmetropic) – refractive error would reveal the possible candidate variants that may account for the trait in affected (in this case severely shortsighted) group only. Since this study is aiming to disclose the genetic background of high myopia, refractive error in all association analyses reported in this thesis was treated as a dichotomous phenotype.

Refractive error shows a leptokurtotic distribution that is skewed towards myopia: the errors are clustered near zero and their prevalence falls exponentially moving away from zero in either direction (Figure 2.4). Based on the bimodality of the myopic limb of the refractive error curve (Figure 2.4), the criteria used to determine affectation status was historically set to be more or equal to -6.0 Dioptre [271]. Several genetic studies on high myopia have adopted this threshold [172, 175, 176, 180] and, therefore, analyses reported in this thesis were performed using this criterion.

The distribution of refractive error changes with age. The average refractive error of newborns is around +3.0 dioptres of hyperopia, shifting towards +1.0 dioptres by one year of age [272]. Based on studies of mostly Caucasian children, myopia typically appears between 6 and 12 years of age [48, 273, 274]. If an association study examines refractive error in form of a dichotomous variable, determining affectation status as high myopia (in this case more or equal to -6.0 Dioptres), it is important to take into account the age of the participants as young children may develop severe shortsightedness after their refraction was acquired for the analyses. The Family Study of Myopia recruits mostly adults over 18 years of age. There are no children younger than 6 years of age and only one participant was aged 12 years of age at the time the association analyses of this study were performed.

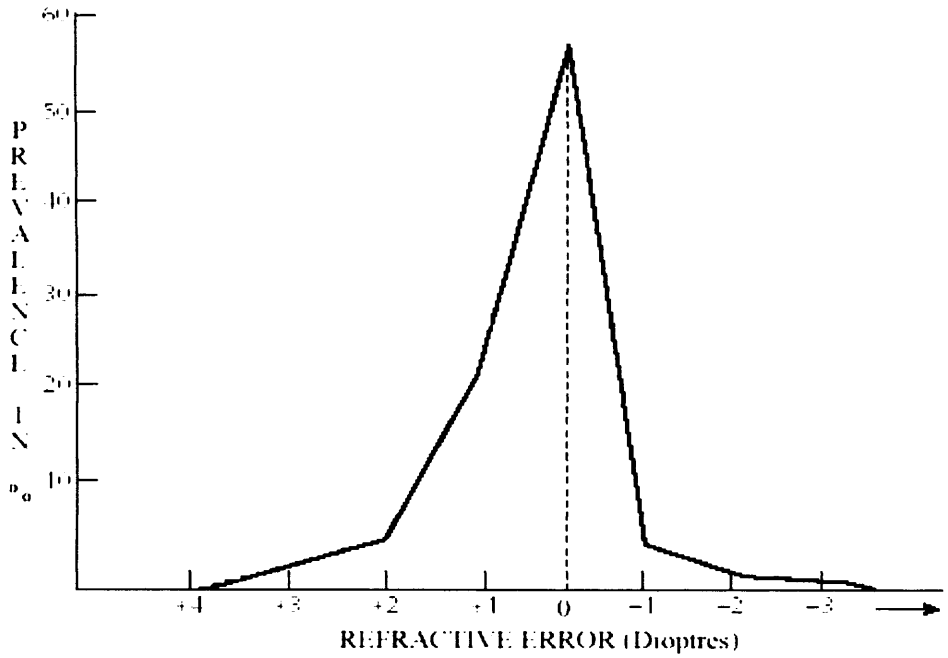


Figure 2.4 The Distribution Curve of Refractive Error
(Taken from [271]).

2.4.4.2 The Statistic of an Association Test

The statistic of an association test can vary depending on whether it is a family-based or case/control study. Most analyses are either likelihood based ratio/score tests or χ^2 / Fisher tests of independence.

Family-based association studies originally involved a so-called transmission disequilibrium test (TDT) only [135]. This method can be considered to be a version of a χ^2 test of independence: the test statistic is calculated based on 2 x 2 table constructed with counts of alleles that are transmitted or non-transmitted to affected offspring in a small trio (mother, father and child) family. However, the original TDT test has several limitations (section 1.3.4.1).

Many family-based association tests are based on the regression model: $Y = \mu + \beta_g g + \beta_x x$, where Y is the observed trait, μ is the population mean, β_g is the additive effect for

each allele, g is a genotype score, β_x is a vector of covariate effects and x reflects the covariate status of each subject. To test for association, a multivariate likelihood test can be performed. First the likelihood is maximized under the null hypothesis of no association and under the constraint that β_g is zero (L_0). Then the procedure is repeated without constrains on parameters to obtain L_1. A likelihood ratio test (LRT) statistic to evaluate the evidence of association would then be equal to $2\ln L_1 - 2\ln L_0$. Such an LRT statistic requires that L_0 and L_1 be maximized for each SNP. That can become computationally challenging on a genome-wide scale [275].

An alternative approach is to first fit a simple model without β_g for each family and then to calculate a so-called score statistic, where, along with the expected genotype g vector determined based on the available marker data, an additional E(g) vector with identical elements is created to give the unconditional expectation of each genotype score. [275].

Both LRT and score statistics asymptotically follow a chi-square distribution with one degree of freedom. However, it is important to note that the distribution of the score statistic will deviate from χ^2 when a linked major gene effect is large. Therefore, score statistics should be used for an initial phase of genome-wide analysis and LRT should be used for re-evaluation of statistical findings in screening steps, to avoid an excess of false positive results in the regions of strong linkage [275].

Although it is possible to restrict analyses to complete data only, such an approach would result in loss of power in comparison to one that would accommodate it [147]. Missing data can originate because the recruitment of both parents is not always feasible in practice or because of failed/incorrect genotyping. To overcome the missingness of parents, information from siblings can be used if available [148, 276]. Alternatively, maximum likelihood can offer solution for any kind of missing data [147, 277, 278].

Case/control association analyses are usually based on tests for independence (χ^2 or Fisher test) at both allelic and genotypic levels as well as on calculation of odds ratios and relative risks. In addition, logistic (binary trait) or linear regression (quantitative trait) can also be applied [127].

2.4.5 Multiple Testing and its Correction

Association studies involving many markers give rise to the problem of multiple testing, which results in an increased number of false positives, thus necessitating a correction in the nominal significance level. In genetic studies the risk of false discovery is very high because only few among all tested markers will have an effect in the case of a complex disease. Indeed, it has been speculated that out of 20 reported association studies, 19 are false [279].

Typically, a test is declared significant if the calculated p-value is less than a chosen threshold value. A type I error is the situation of rejecting the null hypothesis when it is true. This produces a false discovery or a false positive result.

The traditional approach for controlling false discoveries is to maintain a desired probability that a study produces no more false positives than a specified error rate. This method controls the error rate for the whole set of tests (e.g. genome-wide tests) [280]. One of the most well known procedures in this group is the Bonferroni adjustment: the cut off p-value is divided by the number of tests performed [281]. This one-step method has, however, been proven to be conservative, leading to loss in the power of a study. In addition, it performs well only when all markers are independent [280]. If more than one marker has an effect, a step-wise procedure is preferred. The main idea is that if one of the null hypotheses is rejected, it cannot be considered true anymore. Therefore, the correction can be made to the number of tests minus one, rather than the number of tests as such [282].

Rather than focusing on the risk of false discoveries, it can be argued that it may be better to calculate the ratio of false positives: the probability that a randomly selected marker among significant ones is false [283]. This ratio is called the false discovery rate (FDR) and is fundamentally different from traditional approaches. Firstly, because the risk of false discoveries in genome-wide analyses is high, the traditional correction will heavily penalize test statistics by imposing very small threshold p-values. However, a large association study is also likely to discover more true positives. Thus, FDR will reward it by focusing on the proportion of false discoveries divided by all rejected tests (including

false but also true positives). Due to their small effect sizes, the power to detect genes responsible for complex diseases is already low, thus, it may be more advantageous to allow an occasional false discovery to improve the chances of finding an effect instead of further sacrificing power. Furthermore, because there will be multiple significant genes with small effects, the consequences of false discovery may not be as severe as in a single gene analyses where a discovery implies a strong claim that the cause has been found [280].

A second important difference is that in contrast to traditional methods, FDR does not concentrate on the number of tests performed. Instead, it is based on the so-called p_0 – value: the proportion of markers with no effect on a disease or, in other words, the probability that a randomly selected marker has no effect. The higher this proportion the more likely it is that a discovery is false [280]. This provides a better basis for comparison of different studies (e.g. replication analyses): the number of tests performed by different researchers is arbitrary and may depend on such factors as budget or genotyping capacity, whereas parameter p_0 is not arbitrary and applies similar standards to different studies [280].

It is noteworthy, however, that FDR, as any method, has its disadvantages as well. The major limitation is that the p_0 –value and effect size are unknown. The p_0 –value commonly is assumed to be one, which results in a conservative test because the high p_0 will produce a low threshold value. To avoid this bias, p_0 can also be estimated from the data [284].

Finally, correction for multiple testing can also be done by the use of permutation [285]. Random permutations of the data are obtained by sampling from the same set of observations without a replacement. The limitation of this procedure is that the number of permutations should be large enough (preferable exceed 100/threshold p-value) to be able to estimate low p-values.

CHAPTER III.

QUANTITY AND QUALITY OF DNA EXTRACTED FROM MOUTHWASHES

In large-scale genetic linkage and association studies there is a need for a cost-effective, safe and efficient method of obtaining DNA. An attractive approach is to use buccal cells as, in comparison to blood, they offer a non-invasive and more easily collected source of cellular material. Various methods of buccal cell collection have been proposed, such as mouthwash, cytobrush and type cards [286, 287]. Among these procedures mouthwash can be performed by study participants without supervision, has the advantage of being collected via mail [288, 289] and yields the highest amount of DNA [286].

Despite these numerous advantages, there is a need for caution in using DNA extracted from buccal cells because of the presence of non-human DNA in mouthwash samples, e.g. from oral bacteria or food remnants. Once the DNA is extracted, the biggest issue is to distinguish between different origins of DNA and accurately estimate the quantity/quality of human DNA.

The following three sections describe three experiments performed in order to gain insight into quantification and quality assessment of mouthwash-derived DNA.

3.1 Experiment 1. Quantification of Mouthwash-extracted, Human DNA

3.1.1 Introduction

Correct DNA quantification is essential for many genetic applications, e.g. efficient high-throughput genotyping and sample conservation. Inaccuracy in DNA quantification can result in the unnecessary consumption of DNA [290], can lead to lower confidence in scoring genotype by increasing the variability in the amount of PCR product used by most genotyping technologies [291], and can give rise to wrong allele frequency estimations when such samples are pooled [292].

Conservation of original DNA samples is important to validate previous studies and to allow for future studies, representing a critical goal for the efficient utilization of research resources [290].

In light of the above, an accurate and reliable method for DNA quantification is essential for any genetic study. Several methods have been developed to quantify DNA: basic UV spectrometry, gel-based techniques, fluorometry and amplification (PCR). Early techniques (spectrometry, gel electrophoresis, fluorometry) simply measured total DNA, but newer PCR can specifically measure human DNA. In addition, spectrophotometry usually overestimates the amount of human DNA not only because of the presence of non-human DNA, but also because its measurements are influenced by UV-absorbing contaminants like proteins and phenol, which may interfere with the quantification results [293, 294]. In addition, the DNA concentration must be at least 3ng/μl in order to give reliable results with UV spectrophotometry [295]. Quantification of human DNA can also be inaccurate when estimated by DNA-specific dye based fluorometry or gel electrophoresis as a fluorophore dye cannot distinguish between DNA of different origins [294].

Performing conventional PCR for DNA quantification is time-consuming, while real-time reaction requires an expensive machine and reagents that may not be available in every laboratory. This experiment, thus, assessed classic and newer DNA quantification techniques and compared their performance to that of qPCR.

3.1.2 Materials and Methods

3.1.2.1 Subjects and DNA samples

Subjects who took part in this experiment were volunteers from the School of Optometry and Vision Sciences. Each of the participants was asked to provide three mouthwashes.

3.1.2.2 Spectrophotometry

The principal of spectrophotometry is described in section 2.2.2.1. To calibrate the spectrophotometer, TNE (10mM Tris, 100 mM sodium chloride (NaCl), 1 mM EDTA, pH = 8.0) was used as reference (blank) and calf thymus DNA solution of known concentration (50 μg/ml) was used as a standard. Samples were diluted in TNE 1:100 in triplicate, i.e. every sample was measured 3 times. The ratio of A_{260}/A_{280} for each

assessment was determined and those with less than 1.79 or greater then 2.10 were rejected. The concentration of DNA in a sample was ascertained as described in the section 2.2.2.1.

3.1.2.3 Fluorometry

The principle of fluorometry is described in section 2.2.2.2.

A dilution of stock SYBR-green in TE (10 mM Tris, 1mM EDTA) 1:10 000 was prepared. A linear range of standard, human placenta DNA (Sigma) dilutions from 1 to 2 μg/μl were used as standards. Two micro litres of samples and prepared standards were diluted in 2000 micro litres of SYBR-green-TE (1:1000). Dilutions were produced in triplicates. A mixture of dye and TE was measured without any DNA as a blank.

The instrument settings were set according to the results of previously carried out tests with calf thymus DNA of known concentration. The wavelengths of excitation and emission were taken from the SYBR-green manufacturer's package inserts and were verified from the literature [221]. The fluorimeter was set as follows: exciting wavelength = 497 nm, emission wavelength = from 500 to 540 nm, scan speed = 240 nm, the excitation monochromators were adjusted to a band width of 5 nm and the emission's to 2.5 nm.

Using Microsoft Excel, a standard curve for samples with known concentration and their units of fluorescence was established by performing linear regression (section 2.4.1.2). The equations of regressions for each measurement were used to determine the concentration of DNA in the test samples.

3.1.2.4 Ultraviolet Transilluminator Gel Imaging System

The agarose gel electrophoresis procedure is described in section 2.2.2.3.1.

One percent agarose gels were run using boric acid-sodium hydroxide buffer. Samples for agarose gel electrophoresis were prepared in the following way: 1 μl of a purified, mouthwash-derived DNA sample was mixed with 2.4 μl of SYBR-green-ficoll (15%

Ficoll 400, 0.5% xylene cyanol FF, 10mM EDTA, 1:50 dilution of stock SYBR Green I (Molecular Probes Ltd, Paisley, UK) solution) and 8.6 µl of water. Ten microliters of this mixture were loaded into the gel.

Calf thymus DNA was used as a standard. A calibration curve was prepared with a linear range of dilutions from 0.5 to 2 µg/µl. Electrophoresis was performed at a voltage of ~60 Volts for 40 minutes. Every sample and standard was run 4 times (two gels with duplicate samples and standards). As a control, one well of the gel was loaded only with dye and water and no DNA. A DOC-008.XD (UVItec Ltd, Cambridge, UK) camera system coupled to an ultraviolet transilluminator was used to take a digital photograph of the gel and Quantity One software package was applied to determine the density of the DNA fragments. A standard curve was constructed (amount of DNA in standards versus their density) in Windows Excel. Linear regression analyses were performed for each gel and regression equations were adopted to calculate the amount of DNA loaded in each well of each gel.

3.1.2.5 Quantitative Polymerase Chain Reaction (qPCR)

For this experiment, conventional PCR amplification of microsatellite marker D7S3056 (forward 5' CAA TAG CCC TGA CCT TAT GC, reverse 5' TAC CTA CCT ACC TAC CTC TAT GGC) was carried out. The principle of PCR is described in section 2.2.2.4.

Triplicate dilutions of each DNA sample were prepared to achieve ~5ng/µl concentration. Human placenta DNA was used as a standard and a linear range of concentrations from 1 to 5 ng/µl was also prepared. Essential reagents for PCR were mixed to reach the following final concentrations of 1x HotStar PCR buffer (Qiagen Ltd, Crawley, UK), 1.5 mM MgCl2, 0.2 µM each dNTP and 1µM of each primer. Each reaction contained 0.5U HotStar Taq polymerase (Qiagen Ltd, Crawley, UK). Initial step of 15 minutes at 95°C served to activate the HotStarTaq polymerase enzyme. Amplification was achieved by 25 cycles of the following steps: denaturation at 94°C for 1 minute, annealing at 60°C for 1 minute and extension at 72°C for 1 minute. Ficoll-EDTA (15% Ficoll 400, 0.5% xylene cyanol FF, 10mM EDTA solution) was added to PCR products, and agarose gel

electrophoresis was performed in a 2% agarose gel at the voltage of ~60 Volts for 40 minutes. To visualize the DNA, ethidium bromide staining was performed. The optical density of gel bands was analyzed as described above.

3.1.2.6 Statistical Analyses

Yields of DNA in each sample, calculated using four methods described above, were tested with ANOVA and compared with Dunnett's test to see whether there was any significant difference between results of qPCR and the three other procedures. Each dataset for each quantification method was weighted by its average. PCR results were treated as a reference group for Dunnett's test.

3.1.3Results

3.1.3.1 Subjects and DNA samples

In total, five subjects took part in the experiment and, thus, 15 DNA samples were available for analysis.

3.1.1.2 DNA Quantification and Statistical Analyses

Each sample was quantified three times with each of the four methods examined: spectrophotometry, fluorometry, gel electrophoresis and qPCR (180 measurements in total; Table 3.1). Out of analyzed five subjects, two had only one acceptable spectrophotometry reading.

Table 3.1. Results (DNA yield) of four Potential Quantification Methods.

Each sample was measured three times. Mouthwash two of subject two and mouthwash two of subject four had only one acceptable spectrophotometry reading (A_{260}/A_{280} less than 1.79 or more than 2.10). SD is the abbreviation for standard deviation.

Subject	Mouthwash	DNA Yield (μg) measured by			
		Spectrophotometry	Fluorometry	Electrophoresis	PCR
1	1	94.50	77.59	66.19	71.08
		87.75	70.83	62.50	51.64
		87.25	72.43	42.80	61.36
Mean (SD)		89.83 (4.05)	73.62 (3.53)	57.16 (12.58)	61.36 (9.72)
1	2	8.50	11.55	7.06	5.94
		7.75	11.55	6.89	3.36
		10.00	11.55	5.61	4.65
Mean (SD)		8.75 (1.15)	11.55 (0.00)	6.52 (0.79)	4.65 (1.29)
1	3	47.25	89.52	60.39	68.18
		45.25	90.07	63.18	70.43
		44.25	84.29	53.71	69.30
Mean (SD)		45.58 (1.53)	87.96 (3.19)	59.09 (4.87)	69.30 (1.13)
2	1	20.50	21.19	47.36	15.37
		17.50	22.32	47.88	18.57
		16.50	20.00	38.38	18.69
Mean (SD)		18.17 (2.08)	21.17 (1.16)	44.54 (5.34)	17.54 (1.88)
2	2	13.00	18.09	39.88	14.08
		N/A	15.29	42.72	12.65
		N/A	16.74	37.29	12.15
Mean (SD)		N/A	16.71 (1.40)	39.96 (2.72)	12.96 (1.00)
2	3	30.75	27.76	19.27	24.71
		29.50	30.24	27.74	32.65
		23.00	25.48	40.48	33.68
Mean (SD)		27.75 (4.16)	27.83 (2.38)	29.16 (10.68)	30.35 (4.91)
3	1	40.50	36.90	40.14	34.12
		49.00	33.25	41.05	34.62
		47.00	23.18	38.52	34.67
Mean (SD)		45.50 (4.44)	31.11 (7.11)	39.90 (1.28)	34.37 (0.25)
3	2	52.75	42.59	40.14	44.12
		51.75	40.15	41.05	32.97
		51.50	37.94	21.87	33.54
Mean (SD)		52.00 (0.66)	40.23 (2.33)	34.35 (10.82)	36.88 (6.28)
3	3	84.00	40.22	58.77	44.12
		81.00	34.03	53.96	22.97
		79.75	32.52	59.05	33.54
Mean (SD)		81.58 (2.18)	35.59 (4.08)	57.26 (2.86)	33.54 (10.58)

Table 3.1 Results (DNA yield) of four Potential Quantification Methods (Continuation)

Subject	Mouthwash	DNA Yield (μg) measured by			
		Spectrophotometry	Fluorometry	Electrophoresis	PCR
4	1	64.25	40.01	50.07	49.96
		62.50	47.08	41.41	28.72
		59.50	36.57	23.60	39.34
Mean (SD)		62.08 (2.40)	41.22 (5.36)	38.36 (13.50)	39.34 (10.62)
4	2	9.50	10.71	19.04	9.42
		N/A	9.78	23.38	16.67
		N/A	10.06	17.39	9.95
Mean (SD)		N/A	10.18 (0.48)	19.94 (3.09)	12.01 (4.04)
4	3	29.00	54.20	25.50	68.43
		21.00	46.50	21.57	74.23
		21.75	63.48	35.50	58.30
Mean (SD)		23.92 (4.42)	54.73 (8.50)	27.52 (7.18)	66.99 (8.06)
5	1	60.00	45.33	37.19	24.75
		55.00	49.02	25.72	41.26
		54.25	46.85	13.16	33.01
Mean (SD)		56.42 (3.13)	47.07 (1.85)	23.56 (12.02)	33.01 (8.26)
5	2	78.50	36.86	36.59	37.99
		76.50	37.45	33.60	36.26
		94.25	49.49	28.16	31.91
Mean (SD)		83.08 (9.72)	41.27 (7.13)	32.78 (4.27)	35.39 (3.13)
5	3	26.50	15.23	17.73	18.16
		27.50	15.45	32.09	38.01
		23.50	15.16	24.91	28.08
Mean (SD)		25.83 (2.08)	15.28 (0.15)	24.91 (7.18)	28.08 (9.93)

Compared to the amount of human DNA measured by qPCR, yields calculated by spectrophotometry, fluorometry and electrophoresis were overestimated by 33.64%, 7.7% and 4.08% respectively.

All groups of readings for each subject had no significant difference in their variability (Levene's test p-value > 0.05, Table 3.2). ANOVA analysis revealed a significant difference between the groups of measurements for all five subjects ($p < 0.05$), except for the third mouthwash of subject two (Table 3.2). Post-Hoc analysis with Dunnett's test showed no significant difference between fluorometry and PCR in 13 samples out of the 15 examined (Table 3.2). This pattern was followed in 3 samples out of 13 for spectrophotometry, and in 9 samples out of 15 for agarose gel electrophoresis (Table 3.2).

Table 3.2 Comparison of human DNA quantification methods
(ANOVA and Dunnett's Tests Results)

ANOVA and Dunnet's post-hoc test were used to test whether the measurements of four examined DNA quantification techniques were different from each other (non-significant result is highlited).
Note that the second mouthwash of subject two and the second mouthwash of subject four had only one acceptable (A_{260}/A_{280} less than 1.79 or more than 2.10) spectrophotometry reading. Thus, Dunnett's test was not performed for those samples. Levene's statistic was calculated to test for homogeneity of variances between groups of measurements.

Subject	Mouthwash	Levene's p-value	ANOVA p-value	Dunnett's Test p-value (PCR versus)		
				Spectrophotometry	Florimetry	Electrophoresis
1	1	0.123	0.001	0.002	0.112	0.772
1	2	0.871	0.000	0.001	0.000	0.078
1	3	0.121	0.000	0.000	0.000	0.000
2	1	0.267	0.000	0.976	0.143	0.000
2	2	0.866	0.000	N/A	0.100	0.000
2	3	0.270	0.961	0.942	0.934	0.977
3	1	0.108	0.010	0.023	0.609	0.284
3	2	0.161	0.013	0.019	0.761	0.841
3	3	0.200	0.000	0.000	0.886	0.001
4	1	0.438	0.013	0.015	0.971	0.986
4	2	0.598	0.001	N/A	0.556	0.002
4	3	0.509	0.000	0.000	0.152	0.000
5	1	0.352	0.003	0.011	0.105	0.425
5	2	0.971	0.000	0.000	0.451	0.872
5	3	0.396	0.025	0.973	0.017	0.719

3.1.4 Discussion

Due to variety in the origins of sample from which DNA can be extracted (e.g. buccal cells or blood) and in the method of purification, low yields of DNA and/or the presence of contaminants are frequently encountered in molecular biology applications. This emphasizes the need and importance of a method capable of quantifying low levels of DNA, with (1) minimal consumption of the total available sample, and (2) minimal influence of sample impurities on its accuracy. Precise quantification of DNA is crucial for efficient molecular procedures, such as genotyping, in order to maximize high-throughput completion rates, accuracy and reproducibility. It also enables good sample management and reduces unnecessary DNA consumption.

In this experiment, the performance of four potential DNA quantification methods was compared. As expected from previous studies [293, 294], the amount of human DNA measured by spectrophotometry, fluorometry and electrophoresis was overestimated compared to the qPCR results that provided human-specific assessment.

Until recently, specrophotometry has been the traditional method of measuring DNA concentration as it does not require complicated equipment or multi step sample preparation. In addition, spectrophotometry has been proved to have small sample-to-sample variability [290]. In this experiment, its coefficient of variability (standard deviation divided by the mean of a sample) was the smallest (8.36% on average), confirming that it is indeed a reproducible method for DNA quantification. Nonetheless, non-specificity to DNA makes spectrophotometry readings not acceptable as a measure of human DNA concentration: the overestimation of spectrophotometry over qPCR was the largest at 33.64%, suggesting that more than 1/3 of what spectrophotometry quanitfies is not actually human DNA. Moreover, spectrophotometry is also the least sensitive method, requiring a relatively large amount of sample DNA for its quantification.

In contrast to spectrophotometry, fluorometry and gel electrophoresis are both DNA specific and sensitive due to DNA-binding fluorophores applied in these techniques. Electrophoresis has an additional advantage of assuring of DNA quality (degraded or non-degraded). However, because gel electrophoresis requires a number of sample preparation

steps, its reproducibility is inferior to that of fluorometry (coefficient of variance: 20.5% versus 8.42%), whose sample preparation is simpler. DNA quantities measured by fluorometry, in this experiment, were not significantly different from those estimated by qPCR for the majority of DNA samples (87% concordance), at least in relative terms. This confirms the previously published results of fluorometry being useful as an accurate alternative to a qPCR for DNA quantification [293, 294]. Nonetheless, it does not give any measurement of purity (like an A_{260}/A_{280} ratio of spectrophotometry), nor does it certify that DNA is not degraded (like gel electrophoresis) and requires more complicated and expensive equipment than either spectrophotometry or electrophoresis.

3.1.5 Conclusion

In light of the above, fluorometry has the potential to substitute for human-specific qPCR, provided that DNA is not degraded and is primarily of human-origin.

3.2 Experiment 2: The quality of mouthwash-extracted, human DNA: effect of lag time between mouthwash rinse and DNA extraction on quality of the mouthwash-derived DNA

3.2.1 Introduction

Buccal cells are an important source of DNA in epidemiological and genetic studies, and, thus, it is essential to determine how to maximize the amount and quality of DNA collected from this cellular material. Among several methods proposed for buccal cell collection [286, 287, 289, 295-297], mouthwash has proved to give the greater DNA yield and to be easier to perform than the others [286, 289]. However, the conditions of mouthwash performing/collection may affect the DNA extracted from it. The circumstances of mouthwash performance such as swish time and tooth brushing/eating/drinking before collection have already been examined [286], showing that for the best results a mouthwash should be performed before tooth brushing, eating or drinking; while swish time (30 seconds versus 1 minute) has no effect on DNA derived from buccal cells [217]. Lag time between mouthwash rinsing and processing, on the other hand, has been proposed as a possible cause of poor DNA quality [217, 286, 296] extracted from buccal cells, but this issue has not been investigated in detail. As the Family Study of Myopia collects mouthwashes by post, in this experiment, the effect of delay of 0-3 days in the time between a mouthwash rinse being carried out and DNA extraction being completed (mimicking the delay that would be experienced by posted mouthwash samples) to find out whether such a time-delay had an effect on DNA degradation evaluated by agarose gel electrophoresis. In addition, the extend of association between DNA degradation and a subject was also assessed.

3.2.2 Materials and Methods

3.2.2.1 Subjects and DNA samples

Ethical approval for this experiment was granted by the Cardiff University Human Sciences Research Ethics Committee. The experiment adhered to the tenants of the Declaration of Helsinki and all participants provided written informed consent.

Subjects of this experiment were university students, who each provided one mouthwash per day on 12 separate days. These volunteers brought their mouthwash samples to the laboratory on the day they were obtained. The single mouthwashes collected each day were assigned to one of the four groups A-D (with 3 mouthwashes per group), with samples being stored at room temperature for (A) zero, (B) one, (C) two, or (D) three days before being processed.

DNA was extracted as described in section 2.2.1.

3.2.2.2 UV Transilluminator Gel Imaging System

DNA degradation was examined with agarose gel electrophoresis as described in section 3.1.2.4 of the previous experiment. Electrophoresed DNA was visualized using a digital imaging system.

DNA degradation appeared as an "all-or-nothing" phenomenon, making the effect straightforward to score by eye, and because existing automated image analysis systems for electrophoresis gels are not designed to score DNA degradation, it would have necessitated the development of a custom software to perform the task. Since (1) the time taken to score the gels by eye was shorter than the time required for software development, and (2) an automated system seemed unlikely to provide a greatly improved level of reproducibility over scoring by eye, it was decided that DNA degradation would be scored by visual inspection: samples showing a smear instead of a well-defined band were recorded as degraded.

3.2.2.3 Statistical Analysis

To explore whether a delay prior to DNA extraction (0-3 days) influenced the proportion of DNA samples that were scored as being degraded, logistic regression (section 2.4.2.2) was carried out, with time (in days) as a predictor variable.

3.2.3 Results

3.2.3.1 Subjects and DNA samples

In total 6 volunteers agreed to participate in this experiment. All of them provided 12 mouthwashes (72 samples in total) and DNA was successfully extracted after the appropriate 0-3 days delay.

3.2.3.2 Effect of lag time on DNA degradation assessed by gel electrophoresis

There was no obvious relationship between the amount of lag time a mouthwash sample was stored at room temperature (0-3 days) prior to extraction and the presence/absence of DNA degradation. Some subjects had degraded DNA in all samples (Figure 3.1), whilst for other subjects DNA degradation was sporadic (Figure 3.2). Statistically, lag time had no significant effect on DNA degradation in the logistic regression model (Wald test: $Z = 0.052$, $df = 1$, $P = 0.819$).

Figure 3.1 Gel electrophoresis of 12 mouthwashes of one the subjects examined (Degraded DNA in all samples)

The electrophoresed 12 mouthwash-extracted DNA samples were divided into 4 groups according to the number of days (0-3 days) they were left at a room temperature before further processing. DNA in all samples derived from this individual show signs of degradation.

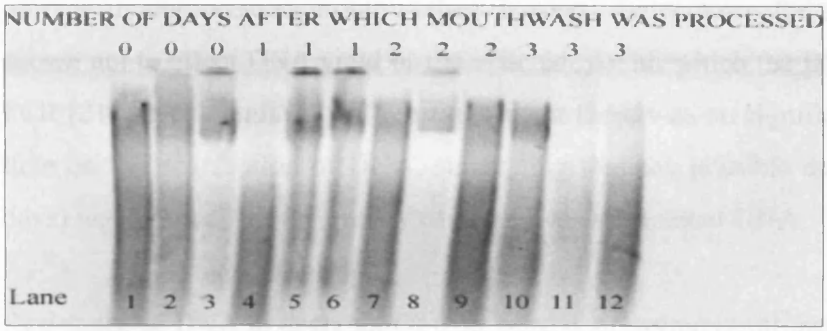
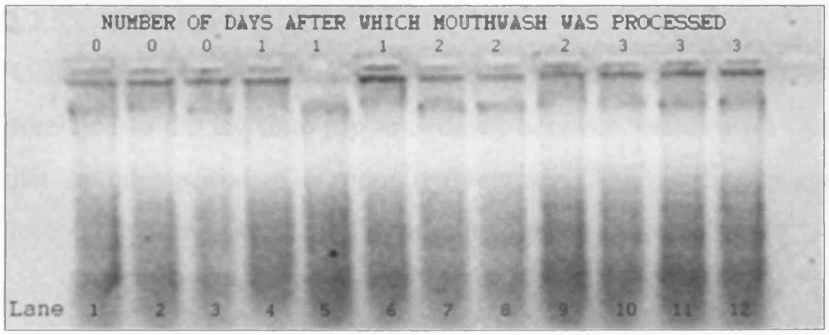


Figure 3.2 Gel electrophoresis of 12 mouthwashes of one the subjects examined (Degraded DNA in some samples)

The electrophoresed 12 mouthwash-extracted DNA samples were divided into 4 groups according to the number of days (0-3 days) they were left at a room temperature before further processing. DNA is degraded in some samples (for example: lane 5) derived from this individual, but not in others (for example: lines 6).



3.2.4 Discussion

The effect of lag time between mouthwash rinsing and processing on the degradation of extracted DNA was examined in this experiment, as it has been suggested to be one of the potential causes for poor DNA quality derived from buccal cells [217, 286, 296].

Storage of unprocessed mouthwashes at room temperature for up to 1 week has been shown not to affect DNA yield or the efficiency with which the DNA can be amplified by PCR [217, 296]. Similarly, in this experiment there was no significant influence of the lag time on the degradation of DNA, suggesting that the possible delay in postage (up to 3 days) would not affect the quality of mouthwash-extracted DNA.

Resistance of DNA to degradation over time is presumably influenced by the composition of the mouthwash solution itself (e.g. the presence or absence of alcohol). In this experiment mouthwashes were performed using sterile saline. DNA has been proved to be stable in saline at room temperature for up to 4 days [298], and, likewise, there was no pattern of increased degradation over 3 days time in this investigation.

3.2.5 Conclusion

Consistent with previous studies, this experiment showed DNA degradation being unrelated to the lag time (up to 3 days) between mouthwash rinse and DNA extraction, that samples spend at a room temperature. This finding suggests that mailing is an acceptable form of collection of mouthwash buccal cells.

3.3 Experiment 3: Quality Assessment of Mouthwash-extracted DNA

3.3.1 Introduction

Buccal cell samples provide a valuable source of human DNA for genetic polymorphism analysis in molecular studies, especially if blood collection is not feasible (e.g. large "field" epidemiological study). The success of PCR reactions using mouthwash-derived DNA has been shown to be dependant on the size of the amplified fragment – the amplification of long DNA fragments being more difficult presumably because of nucleic acid degradation [217, 299]. In addition, mouthwash-derived DNA shows a higher discordance rate than blood-derived DNA in genotyping and whole genome amplification [300].

Together, the above results suggest that buccal cell DNA is inferior to that obtained from blood. To address this issue in greater detail, this experiment evaluated the performance of gel electrophoresis and qPCR as DNA quality control procedures. After checking for degradation with agarose gel electrophoresis, "human-specific" qPCR was carried out to establish whether sufficient DNA of human-origin was present to perform efficient PCR (and to investigate whether degradation noticed with gel electrophoresis had any adverse effect on the outcome of qPCR). Finally, those samples that were judged to be non-degraded by gel electrophoresis and to contain a supra threshold level of human DNA by qPCR were genotyped with Illumina 6k human bead array assay to assess if the quality control techniques were able to identify "high quality DNA" samples.

3.3.2. Materials and Methods

3.3.2.1 Subjects and DNA samples

Recruitment of subjects, collection of their mouthwash samples and DNA extraction procedures are described in sections 2.1 and 2.2.1.

3.3.2.2 UV Transilluminator Gel Imaging System

The quality and quantity of mouthwash-extracted DNA was assessed by gel electrophoresis and gel imaging (section 2.2.2.3). Briefly, after extraction from a mouthwash, DNA was diluted 1:10 in TE and 10 µl of the dilution was electrophoresed on a 1% agarose gel with SYBR green I as a fluorophore.

DNA degradation appeared as an "all-or-nothing" phenomenon, making the effect straightforward to score by eye, and because existing automated image analysis systems for electrophoresis gels are not designed to score DNA degradation, it would have necessitated the development of a custom software to perform the task. Since (1) the time taken to score the gels by eye was shorter than the time required for software development, and (2) an automated system seemed unlikely to provide a greatly improved level of reproducibility over scoring by eye, it was decided that DNA degradation would be scored by visual inspection: samples showing a smear instead of a well-defined band were recorded as degraded.

3.3.2.3 Quantitative Polymerase Chain Reaction (qPCR)

"Human-specific" qPCR was also carried out to provide insight into the likely consequences of this degradation for downstream applications. A measure of the human DNA content of mouthwash-derived DNA samples was obtained in one of two ways: a conventional qPCR reaction followed by agarose gel and scanning densitometry, or a real-time qPCR reaction (section 2.2.2.4.2).

For the standard reaction, samples were diluted with a known volume of TE to give an expected final DNA concentration in the range 1-5ng/µl. The quantitative PCR reaction was performed as described in section 3.1.2.5.

For real-time qPCR, amplification was carried out using a Rotor-Gene 6000 thermal cycler, with SYBR-Green I (Molecular Probes-Invitrogen Ltd, Paisley, UK) as the fluorophore. Quantification of DNA was achieved by constructing a standard curve of calculated Ct versus concentration for a set of DNA standards that were included in each

run (section 2.2.2.4.3). Reaction reagents were mixed to give final concentrations of 1.2 x HotStar PCR buffer (Qiagen Ltd, Crawley, UK), 3 mM $MgCl_2$, 0.24mM dNTPs mix, 1.2 μM of each primer (D7S3056) and 1:40 000 SYBR Green I. Each 10 μl reaction contained 1U HotStar Taq polymerase (Qiagen Ltd) and mouthwash-extracted DNA diluted with a known volume of TE to an expected concentration of 0.5 – 2.5ng/μl. Amplification was achieved using 40 cycles of PCR (denaturation at 94°C for 1 minute, annealing at 60°C for 1 minute and extension at 72°C for 1 minute) after a preliminary step of 10 minutes at 95°C to activate the enzyme.

3.3.2.4 High-throughput SNP Array Genotyping

Mouthwash DNA from those participants that proved to contain sufficient human DNA to provide robust amplification of the test amplifier (D7S3056) and that were scored as non-degraded by gel electrophoresis was sent to the Centre of Inherited Disease Research (CIDR) for genotyping on the Illumina 6k Human bead array [301]. Details of the genotyping procedures of CIDR are available at http:/www.cidr.jhmi.edu/human_snp.html. The proportion of mouthwash-derived DNA samples that were successfully genotyped by CIDR was compared to the results of blood-derived DNA sent at the same time. Genotyping was deemed successful if the sample passed the quality control assessment carried out by CIDR. This was based on the use of Illumina's BeadStudio software GenCall (GC) score (a GC score ranges from 0 to 1 and reflects the proximity within a cluster plot of intensities of that genotype to the centroid of the nearest cluster). All genotypes with GC score below 0.25 were considered as failures. DNA samples with >4% genotyping failures were judged as failed samples.

3.3.2.5 Statistical Analyses

Since DNA degradation appeared as an all-or-nothing event as judged from agarose gels, mouthwashes were scored as either intact or degraded using a binary code. Fisher's exact test and odds ratio (sections 2.4.2.2 and 2.4.2.3) were calculated for a 2x2 table containing counts of the number of first and second degraded DNA samples from the 2 consecutive mouthwashes provided by each subject.

Analysis of qPCR results were based on binary coding as well: samples were scored as having "passed" or "failed" to amplify efficiently, depending on whether they reached a threshold level (this threshold being chosen as representative of the minimum level of PCR product required for successful microsatellite genotyping). Fisher's exact test and odds ratio (sections 2.4.2.2 and 2.4.2.3) were computed for a 2x2 table comprising the number of successful qPCR reactions when template DNA was or was not degraded.

3.3.3 Results

3.3.3.1 Subjects and DNA samples

In total 500 subjects (1000 mouthwashes) were collected for this experiment, as a part of The Family Study of Myopia. DNA was successfully extracted from all mouthwashes and analyzed by gel electrophoresis and by qPCR.

3.3.3.2 Quality of DNA and Statistical Analyses

Degradation was observed in a proportion of samples, evident as a broad smear of fluorescence in place of the usual single, sharp, high molecular weight band (Figure 3.3).

The frequency of DNA sample degradation was 8.9% (95% CI: 7.1-10.7%; N = 1000). Among 52 subjects with degraded first mouthwashes, 37 second samples (71%) also contained degraded DNA (Table 3.3). The odds ratio for DNA degradation in the second sample given degradation of the first sample was 3.13 (95% CI: 1.22 – 7.39), which was statistically significant (P=0.009, Fisher's exact test).

Each DNA sample was also assessed using a qPCR assay with primers targeting a human microsatellite marker D7S3056. Figure 3.4 shows the amplification curve obtained for this qPCR reaction, while Figure 3.5 reflects the specificity of the reaction presenting the melting curve with one expected product, which was confirmed by agarose gel

electrophoresis. The average efficiency of the real-time PCR reaction was 94.8%, while the average r^2 value for the linear regression was 98%.

For the 1000 mouthwash-derived DNA samples tested in total, 85.4% of degraded samples passed the qPCR test, compared with 87.8% of non-degraded samples. Statistical analysis suggested that PCR amplification of degraded samples did not differ significantly from that of non-degraded ones (P = 0.5; Fisher's exact test; Table 3.4). The presence of at least some high molecular weight DNA by gel electrophoresis was associated with successful qPCR amplification (Figure 3.6), although this was not investigated in detail.

Figure 3.3 Gel Electrophoresis of Mouthwash-extracted DNA

Agarose gel electrophoresis of DNA extracted from mouthwashes of 8 subjects (2 samples per subject). Subjects are identified by the figures above lanes. DNA extracted from one of the mouthwashes provided by subject 4 and both mouthwashes provided by subject 5 was found to be degraded.

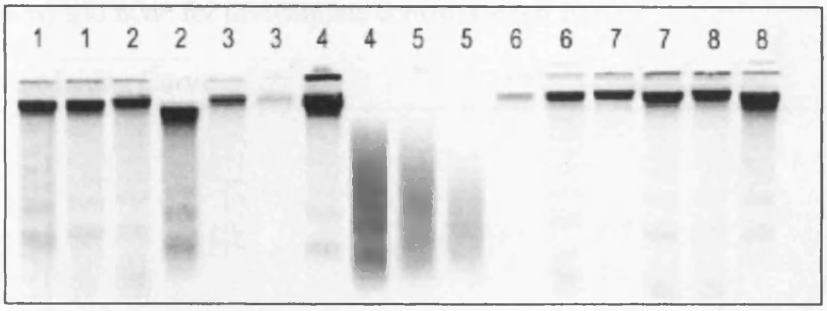


Table 3.3 DNA Degradation in a Subject's First Mouthwash Sample when Analyzed as a Risk-factor for DNA Degradation in their Second Mouthwashes.

		DNA degraded in 2nd mouthwash?		Total
		Yes	No	
DNA degraded in 1st mouthwash?	Yes	9	43	52
	No	28	420	448
Total		37	463	500

Figure 3.4 Real-time PCR Amplification Curve

Amplification of two DNA samples (yellow and purple lines) along with no-template-control (green line) is shown. Amplification starts its exponential growth at cycle number 20 and reaches its plateau phase at cycle number 28.

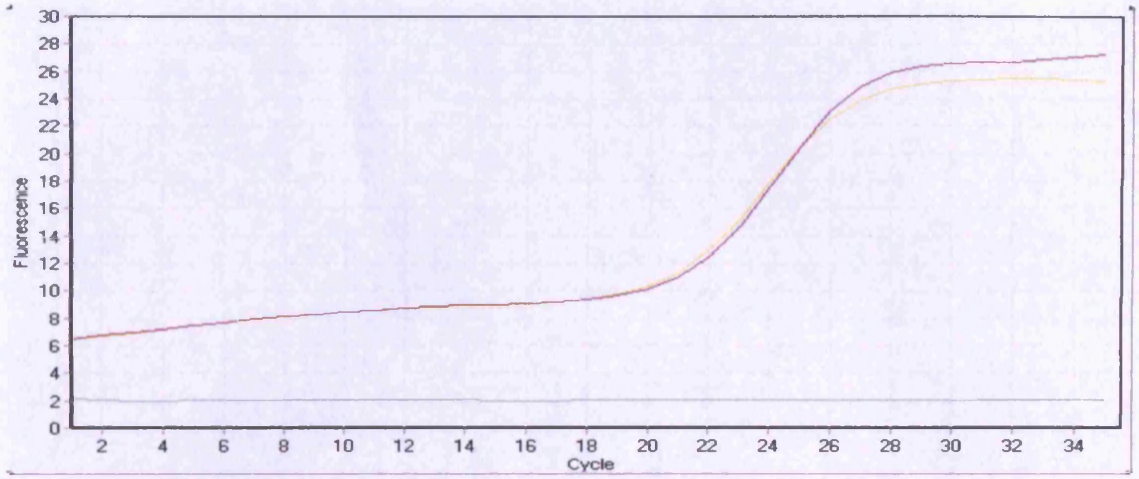
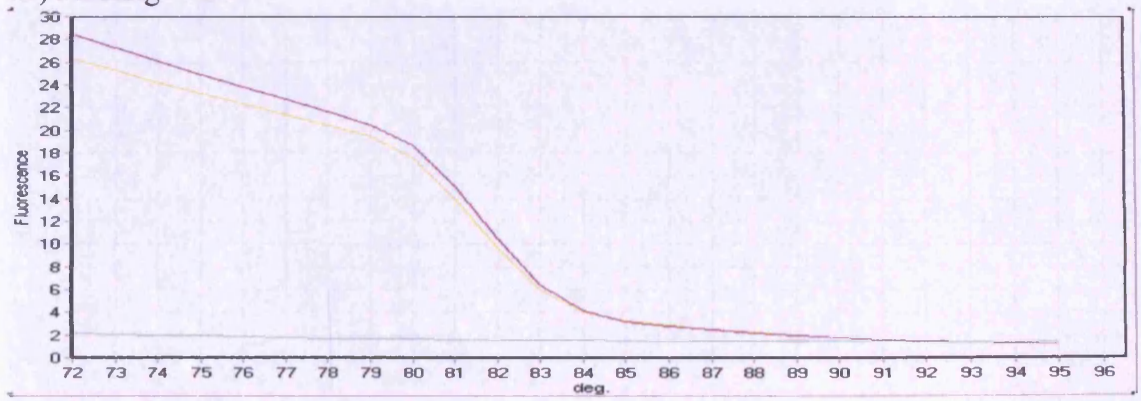


Figure 3.5 Real-time PCR Specificity

An indication of the specificity of real-time PCR was the melting curve: the melting curve of two DNA samples (A.) shows one specific melting point (yellow and purple lines), while the no-template-control has no product (green line). Normalized melting curve (B.) reveals the melting temperature of the product: 82° C for DNA samples (yellow and purple lines) and none for no-template control (green line).

A.) Melting Curve



B.) Normalized Melting Curve

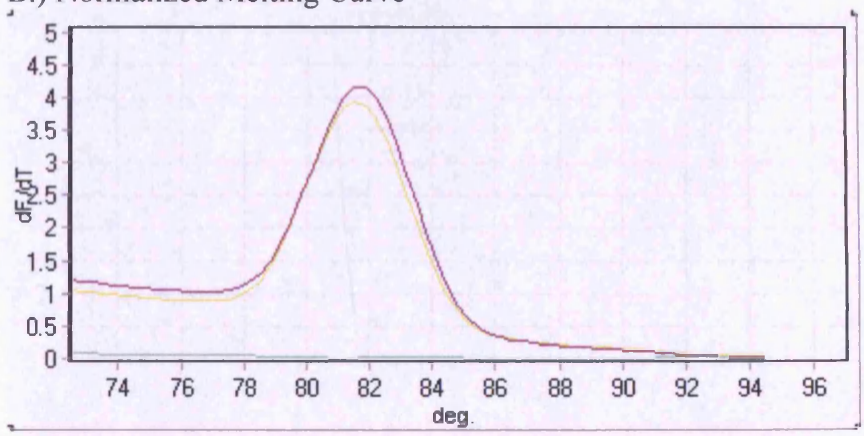
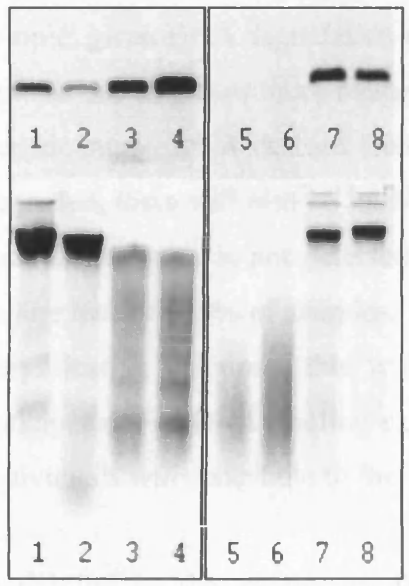


Table 3.4 PCR Success in Degraded DNA Samples

		Successful PCR?		Total
		Yes	No	
Degraded DNA?	Yes	76	13	89
	No	800	111	911
Total		876	124	1000

Figure 3.6 PCR Efficiency of Degraded DNA Samples

Eight mouthwash DNA samples (lower panels) were used as templates for PCR amplification (upper panels). Partially degraded DNA samples containing residual high molecular weight DNA typically permitted efficient PCR amplification (lanes 3 and 4). Severely degraded DNA typically failed to PCR amplify (lanes 5 and 6).



⇦ PCR product

⇦ Non-degraded genomic DNA

3.3.3.3 High-throughput SNP Array Genotyping

Two hundred fifty three mouthwashes proved to provide sufficient human DNA for efficient qPCR and that was judged as non-degraded. These samples were genotyped using the Illumina 6k SNP array platform. There was only one sample (0.4%) that could not be genotyped. For DNA extracted from blood and sent for genotyping at the same time, between 0.6 - 5.3% DNA samples could not be genotyped by CIDR. For the 252 buccal DNA samples that were successfully genotyped, the average number of SNPs that could be genotyped for each subject was 99.7% of the total, and the reliability of SNP genotyping "blind" duplicate mouthwash DNA samples was similarly high (>99.9% concordance).

3.3.4 Discussion

A major finding from this study was the discovery that ~ 10% of DNA samples obtained from saline mouthwashes contained degraded DNA. Furthermore, there was an approximately 3-fold increased risk of DNA degradation in a subject's second mouthwash sample, given DNA degradation in their first. This finding suggests that DNA degradation may be due to one or more factors specific to individual subjects. Therefore, although in a genetic study, DNA derived from the majority of participants (~90%) can score as non-degraded, there will also be individuals (~10%), whose DNA may always be degraded. If such degradation is not detected prior to downstream analyses, it is likely to lead to a failure rate of ~10% of samples. For high-throughput SNP genotyping and whole genome amplification reactions, this will result in reduced statistical power compared to that anticipated. For DNA pooling experiments it may lead to suboptimal results, since fewer individuals will contribute to the genotyping signals that expected.

Variability in the quality of DNA obtained from mouthwashes could arise due to dissimilarity in each individual's oral flora, dietary or lifestyle habits, differences in desquamation of oral mucosa [299] or because of other reasons, such as how exactly the mouthwash rinsing protocol was performed, the composition of the mouthwash solution, and the lag time between mouthwash rinsing and processing. There is a highly diverse and

subject-specific, bacterial flora in the healthy oral cavity [302, 303] that can be affected by smoking [304, 305] and diet [306], and which in turn can lead to DNA damage [307].

The way in which the mouthwash rinsing procedure is performed has been shown to significantly affect DNA yield [298, 308]. Furthermore, cells recovered in mouthwashes are likely to be superficial ones in the process of apoptosis: about 30 % of buccal cells collected from persons with healthy, non-inflamatory oral mucosa show apoptotic signs [309]. Therefore, DNA from certain individuals may be more prone to the signs of DNA degradation noted here.

The lag time between mouthwash performance and DNA extraction from it can also be a possible cause of DNA degradation. However, this was not the case in this experiment as all mouthwashes were processed on the day of their arrival to our laboratory, and delay in doing so up to 3 days was shown to have no significant effect on DNA quality (section 3.2).

Approximately 12% of the DNA samples examined in this study failed to amplify with qPCR. Interestingly, this failure appeared to be independent of visible DNA degradation, suggesting that factors other than this were to blame [310]. It is likely, that carry-over of contaminating substances from DNA extraction played a major role in failure of qPCR in our mouthwash samples. During purification with phenol-chloroform, poor PCR performance [311] and relative loss of human DNA [308] have been observed. In our experience, re-extraction improved PCR performance in approximately 50% of cases, supporting the possible presence of carry-over inhibitors. Nonetheless, since this study attempted to amplify a relatively small product (~200bp) of D7S3056 microsatellite marker (http://www.cephb.fr/cephdb) and previous studies have shown that amplification of long PCR products (>500 bp) is less successful for partially degraded DNA [217, 299], our results suggest that DNA degradation is likely to be more disruptive for demanding downstream analyses.

High-throughput genotyping of mouthwash-derived DNA samples that scored well with gel electrophoresis and qPCR showed high and reliable performance. Thus, this 2-step

procedure may serve as a quality control for DNA of buccal origin collected in "field" conditions. Nonetheless, it is difficult to conclusively state how reliable the 2-step protocol would be since samples that were scored as "degraded" were not sent for genotyping, and, therefore, no comparison between the genotyping success of "degraded" versus "non-degraded" DNA samples was made.

3.3.5 Conclusion

Approximately 10% of mouthwash samples collected using a standardized protocol in our laboratory exhibited signs of DNA degradation. The phenomenon of DNA degradation was shown to be partially subject-specific, although further work will be required to trace the precise cause(s) involved. Therefore, planning a collection of buccal DNA for a large genome-wide study can be made more balanced and cost-effective by taking into the account that ~10% of the collected DNA samples may not be intact.

CHAPTER IV.

MYOCILIN POLYMORPHISMS AND HIGH MYOPIA IN EUROPEAN SUBJECTS

4.1 Introduction

Several highly penetrant genetic loci for non-syndromic myopia have been mapped (section 1.4). However, none of the causative mutations has yet been found. Candidate gene association studies have led to the identification of a number of high myopia susceptibility genes (section 1.4), including the MYOC gene on chromosome 1. Nonetheless, replication of these findings is necessary in order to separate true positives from false positives.

The MYOC gene is best known for its role in glaucoma. Mutations in MYOC can cause both juvenile-onset and adult-onset open angle glaucoma [312, 313]. The MYOC gene consists of three exons (Figure 4.1), and it has been shown that an upstream stimulatory factor is critical for its basal promoter activity [314].

Myocilin (also known as trabecular meshwork inducible glucocorticoid response or TIGR), the protein product of the MYOC gene, was discovered during studies examining proteins that could be induced upon long-term treatment of human trabecular meshwork cells (TMC) with glucocorticoids [315]. In the human eye, myocilin is highly expressed in the TMC, sclera, ciliary body and iris, with considerably lower amounts in retina and optic nerve head. The secreted protein is present in aqueous humor [314]. Aside from glucocorticoid stimulation, the expression of myocilin in TMC is affected by the transcription protein transforming growth factor β (TGF β), mechanical stretch, basic fibroblast growth factor (bFGF) and oxidative stress [314, 316, 317]. Experimental studies show that mutant myocilin isoforms found in patients with juvenile-onset glaucoma are not secreted, but accumulate in the TMC where they are thought to interfere with cell functions. For example, mutant myocilin disturbs the mitochondrial membrane potential [318]. Despite intensive research efforts, however, the precise role of MYOC mutations in glaucoma is unclear.

In addition to glaucomatous involvement, genetic variants in the MYOC gene have also been implicated in causing susceptibility to high myopia [214, 319]. This involvement would be consistent with the possible link between intraocular pressure (IOP)/glaucoma and high myopia, that has been proposed and examined with mixed results in the

literature.

An elevated frequency of glaucoma has been found in myopes: open-angle glaucoma occurs twice as often in the myopic eye as in the non-myopic one [7, 320]. It has also been shown that the probability of developing glaucoma for eyes without myopia is 1.5%, while for eyes with low or high myopia it is 4.2% and 4.4% respectively [321]. In addition, an increased frequency of myopia in patients with open angle glaucoma has been observed [320-322].

The relationship between high IOP and myopia has been suggested to be mediated through near work: there seems to be a general increase of IOP with accommodation and convergence to close distance [323]. In support of this assumption, an increase in intraocular fluid transfer was found during the emmetropisation ("recovery") of previously form-deprived chicks [324]. Moreover, an elevation in IOP of at least 10 mmHg has been shown to lead to a significant increase in the axial length of the eye [325].

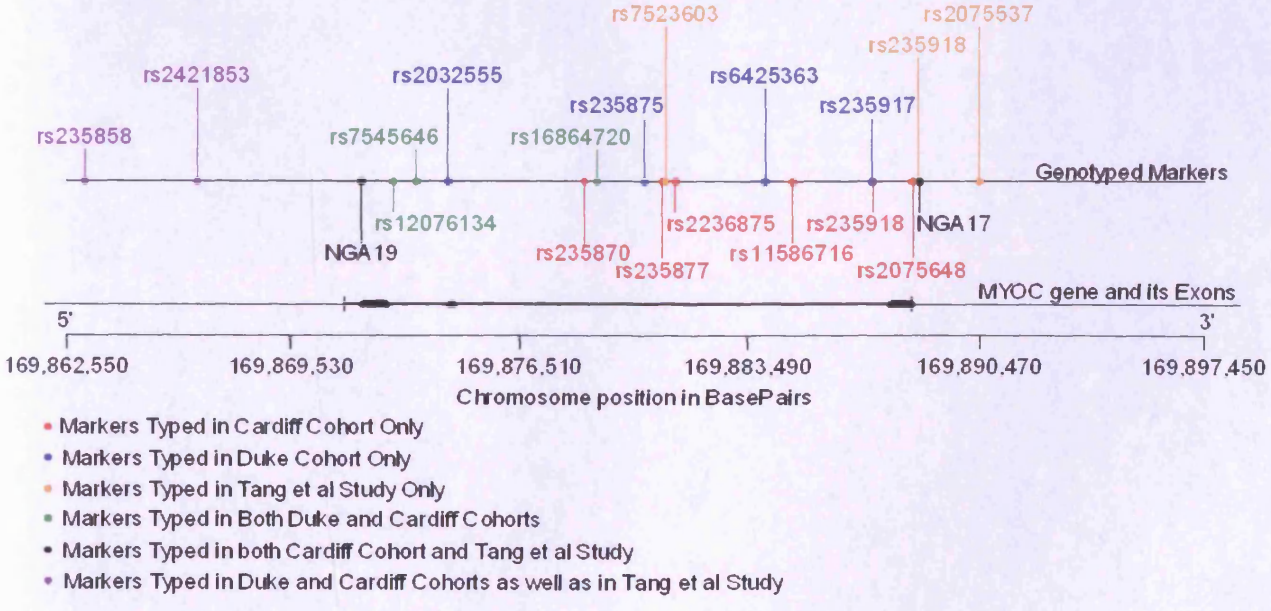
Experiments have shown that the growing eyes of chickens elongate during the day and shorten during the night, which correlates with the IOP circadian rhythm (high during the daytime and low at night) [326]. Nonetheless, there appears to be a phase difference between the rhythms, with the IOP being phase-advanced with respect to the axial length diurnal cycle. Therefore, Nickla et al. [327] proposed that the rhythm in IOP influences ocular elongation in ways other than by simply inflating the eye, for example, by influencing underlying rhythms in sclera extracellular matrix production.

As mentioned above, it has been observed that higher intraocular pressure (IOP) is associated with myopia [7, 274, 322, 328-330]. One interesting study claimed to have success in treating myopia with the IOP-lowering beta-blocker metipranolol [331], but other authors could not replicate this by using a different beta-blocker timolol [332]. It has also been stated that a higher IOP follows the onset of myopia and does not cause it [7, 333]. In addition, a number of studies have failed to establish any correlation between IOP and myopia [334, 335].

It is noteworthy, that some factors that stimulate myocilin expression in TMC have also been implicated in the regulation of postnatal eye growth and myopia, e.g. bFGF, TGFβ and oxidative mitochondrial pathways [186, 210, 336]. In addition, significant genetic linkage has been identified close to the MYOC locus on chromosome 1 in families with myopia from the Beaver Dam Eye Study [183].

Association between MYOC and high myopia was first reported in a case-control study of Chinese subjects from Singapore [319]. An initial attempt to replicate this finding using a similar case-control design in Hong Kong Chinese subjects, however, did not support the association [337]. Later, a larger, family-based association study, also in Chinese subjects from Hong Kong, yielded a significant result [214]. In this latter study, association was found with two microsatellite polymorphisms (NGA17 at the promoter region and NGA19 at the 3' region) and two SNPs (rs2421853, rs235858 at the 3' flanking region). Herein, association between myocilin polymorphisms and high myopia was examined in two independent Caucasian subject groups: a cohort from Cardiff University (UK) and a cohort from Duke University (USA).

Figure 4.1 MYOC gene position, structure and genotyped polymorphisms



4.2.1 Subjects and DNA Samples

Subjects in the UK cohort were recruited as part of The Family Study of Myopia. Participants and their DNA samples were collected as described in sections 2.1 and 2.2.1. Individuals with known syndromic disorders or systemic condition that could predispose to myopia were excluded. All subjects were of Caucasian ethnicity (self-reported "White Europeans").

Subjects in the USA cohort were recruited by the Duke University's Centre for Human Genetics. Genomic DNA was extracted from venous blood using the AutoPure LS® DNA Extractor and PUREGENE™ reagents (Gentra Systems Inc.). All subjects underwent a complete ophthalmic inspection, and individuals with syndromic conditions that could predispose to myopia were excluded. The study was approved by the Institutional review Board at the Duke University Medical Centre. The recruitment and ophthalmic examination of these subjects as well as DNA extraction was performed by Prof. Terri L. Young and her colleagues in Duke University, and did not involve any contribution from me.

4.2.2 Selection and Genotyping of Polymorphisms

The HapMap database lists 25 single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAF) > 5% in the MYOC gene in subjects of European descent. The linkage disequilibrium (LD) structure of the gene in Europeans is shown in Figure 4.2.

In the Cardiff University (UK) cohort, tagging SNPs (sections 1.3.3.2 and 1.3.4.3.2) were selected using the Haploview program [338] conditional on LD (r^2) >0.8 and MAF > 5% (Table 4.2). Genotyping was performed for 12 SNPs within and in the vicinity of MYOC and two microsatellites in the untranslated regions of the gene (NGA17 at 5' and NGA19 at 3' end) including the significant SNPs from Tang et al. study [214]. SNP genotyping was carried out by Kbiosciences Ltd. Microsatellite genotyping was carried out using conventional methods [339]. Briefly, the PCR reaction mixture contained 1x HotStar PCR buffer (Qiagen Ltd), 1.5 mM $MgCl_2$, 200μM each dNTP, 0.3μM of fluorescently-labelled forward primer, 0.3μM of reverse primer, 0.1 U HotStar Taq polymerase (Qiagen Ltd) and

~20ng genomic DNA. Amplification was achieved using 35 cycles of Hot Start PCR (denaturation at 94°C for 1 minute, annealing at 56°C for 1 minute and extension at 72°C for 1 minute) after a preliminary step of 15 minutes at 95°C to activate the enzyme. The primers are shown in Table 4.1. Amplicons were sized using an ABI Prism 310 Genetic Analyzer® (section 2.3.1 in chapter two), run on program D with Genotyper® software (ABI) used to call the alleles.

The selection of tagging SNPs and their genotyping in the USA cohort was performed at Duke University and I was not involved. Tagging SNPs were selected using SNPSelector® conditional on r^2 >0.8 and MAF > 5% in the CEU HapMap population. Genotyping was performed for 9 SNPs, including the significant SNPs from the Tang et al. study [214], using Taqman® allelic discrimination assays (section 2.3.2). SNP tagging and genotyping was carried out by our collaborators in Duke University's Centre for Human Genetics. The SNP selection and genotyping performed by the genetic group in Duke University (USA) did not involve me.

The positions of the SNPs genotyped in this study are shown in Figure 4.1.

Figure 4.2 Linkage Disequilibrium Patterns of MYOC SNPs in European Subjects in the HapMap database

The pairwise correlation of SNPs (based on D`) is shown using red, white and blue squares. Red squares indicate statistically significant (LOD>2) allelic association (linkage disequilibrium, LD) between the pair of SNPs, as measured by the D' statistic; darker colors of red indicate higher values of D', up to a maximum of 1. White squares indicate pairwise D' values of <1 with no statistically significant evidence of LD. Blue squares indicate pairwise D' values of 1 but without statistical significance.
The number in each square is the multi-allelic D' between SNPs.
SNPs are arranged in blocks (depicted here by the thick black triangles) using the "spine of LD" algorithm in Haploview, with adjacent blocks merged if they (1) have multiallelic D' values of at least 0.9, indicating little recombination between blocks and (2) at least 80% of the chromosomes in the resulting merged block are explained by 6 or fewer common haplotypes.

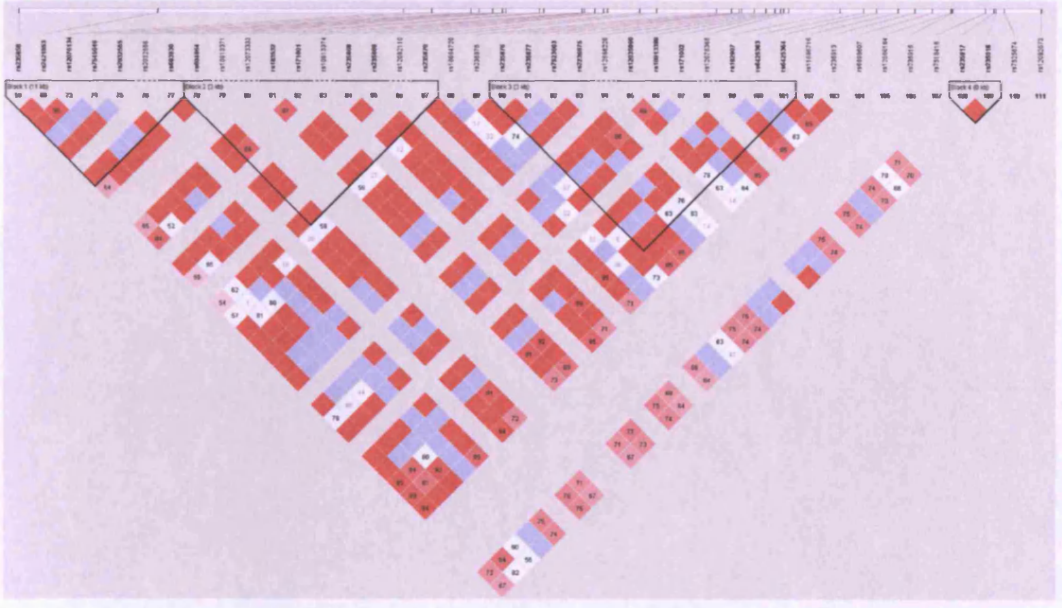


Figure 4.3 Linkage Disequilibrium Patterns of MYOC SNPs Han Chinese Subjects in the HapMap database
(See legend of Figure 4.2 for the explanation of LD diagram)

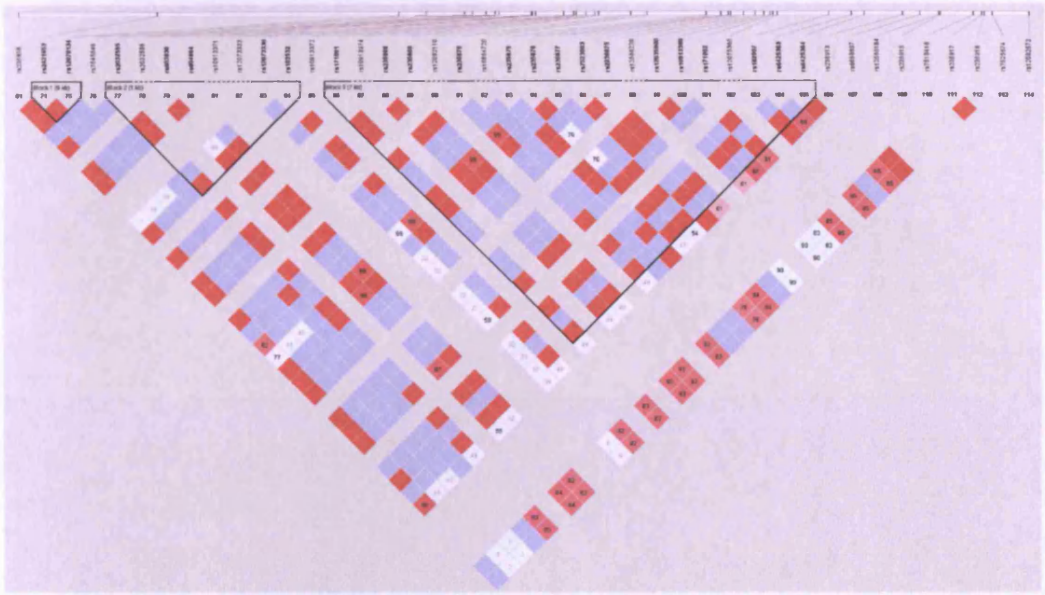


Table 4.1 MYOC Microsatellite Primer Sequences

Primer Name	Primer Sequence
NGA17 forward	GCA CAG TGC AGG TTC TCA A
NGA17 reverse	CCG AGC TCC AGA GAG GTT TA
NGA19 forward	CCA ACC ATC AGG TAA TTC CTT
NGA19 reverse	CCT CAA AAC CAG GCA CAA

Table 4.2 TagSNPs in MYOC (Haploview Results)

A.) Pairwise LD measure for SNPs in MYOC
Only those SNPs are shown that exhibited LD of $r^2 > 0.8$

SNP 1	SNP 2	r^2- value
rs2032555	rs235877	1.0
rs604864	rs235870	0.897
rs183532	rs235870	0.899
rs171001	rs235877	0.89
rs235868	rs235877	1.0
rs235869	rs235870	1.0
rs235876	rs235877	1.0
rs7523603	rs2236875	1.0
rs12035960	rs2236875	1.0
rs171002	rs235877	1.0
rs182907	rs235877	1.0
rs6425364	rs235877	1.0
rs235917	rs235918	0.961
rs2075648	rs2075648	1.0

B.) Chosen Tag SNPs based on Pairwise LD

TagSNP	SNPs captured
rs235877	rs171001, rs182907, rs235876, rs6425364, rs171002, rs2356868, rs2032555, rs235877
rs235870	rs235869, rs604864, rs183532, rs235870
rs2236875	rs12035960, rs7523603, rs2236875
rs235918	rs235917, rs235918
rs235875	Itself
rs11586716	Itself
rs16864720	Itself
rs2075648	Itself
rs7545646	Itself
rs12076134	Itself

Table 4.3 Allele Frequencies of MYOC Markers

Comparison of allele frequencies in the two European cohorts examined in this study and in the Chinese cohort of Tang et al. [214]

A.) *Allele frequencies of microsatellites*

	Cardiff University Cohort			Tang et al
	Family Founders	**Cases**	**Controls**	**Family Founders**
NGA17 alleles				
12 repeats	0.000	0.033	0.028	-
13 repeats	0.597	0.637	0.550	0.501
14 repeats	0.184	0.156	0.170	0.184
15 repeats	0.219	0.174	0.252	0.312
16 repeats	-	-	-	0.003
NGA19 alleles				
11 repeats	-	-	-	0.0015
12 repeats	0.000	0.014	0.000	-
13 repeats	0.342	0.344	0.400	0.218
14 repeats	0.039	0.047	0.004	0.008
15 repeats	0.619	0.595	0.596	0.711
16 repeats	-	-	-	0.060
17 repeats	-	-	-	0.0015

Table 4.3 Allele Frequencies of MYOC Markers (Continuation)

B.) Allele frequencies of SNPs

SNP name	SNP allele	*Cardiff University Cohort*			*Duke University Cohort*			*Tang et al*
		Family Founders	**Cases**	**Controls**	**Family Founders**	**Cases**	**Controls**	**Family Founders**
rs235877	C	0.685	0.655	0.670	-	-	-	-
	T	0.315	0.345	0.330	-	-	-	-
rs235870	A	0.560	0.556	0.551	-	-	-	-
	T	0.440	0.444	0.449	-	-	-	-
rs2236875	G	0.920	0.940	0.930	-	-	-	-
	T	0.080	0.060	0.070	-	-	-	-
rs235918	A	0.353	0.366	0.347	-	-	-	-
	T	0.647	0.634	0.653	-	-	-	-
rs11586716	C	0.264	0.239	0.269	-	-	-	-
	T	0.736	0.761	0.731	-	-	-	-
rs2075648	C	0.869	0.866	0.836	-	-	-	-
	T	0.131	0.134	0.164	-	-	-	-
rs16864720	A	0.131	0.118	0.115	0.121	0.116	-	-
	G	0.869	0.882	0.885	0.879	0.884	-	-
rs7545646	C	0.087	0.074	0.078	0.100	0.116	-	-
	T	0.913	0.926	0.922	0.900	0.884	-	-
rs12076134	G	0.210	0.202	0.232	0.272	0.232	-	-
	T	0.790	0.798	0.768	0.728	0.768	-	-
rs235858	A	0.584	0.590	0.573	0.639	0.607	-	0.600
	G	0.416	0.410	0.427	0.361	0.393	-	0.400
rs2421853	A	0.232	0.217	0.245	0.300	0.277	-	0.270
	G	0.768	0.783	0.755	0.700	0.723	-	0.730
rs6425363	C	-	-	-	0.886	0.900	-	-
	T	-	-	-	0.114	0.100	-	-
rs235917	A	-	-	-	0.284	0.277	-	-
	G	-	-	-	0.716	0.723	-	-
rs235875	C	-	-	-	0.792	0.815	-	-
	T	-	-	-	0.208	0.185	-	-
rs2032555	C	-	-	-	0.239	0.277	-	-
	T	-	-	-	0.761	0.723	-	-

4.2.3 Statistical Analysis

4.2.3.1 Hardy-Weinberg Equilibrium and Mendelian Consistency

The Pedstats package [340] was used to carry out an exact test for Hardy-Weinberg equilibrium (HWE) on unrelated subjects and to check for Mendelian consistency in pedigrees (section 2.3.3).

4.2.3.2 Test for Association (Unphased) and Correction for Multiple Testing

High myopia was examined as a dichotomous trait. Subjects with a spherical equivalent refractive error, averaged between eyes, of < -6.00 D were classified as affected [214]. All other subjects were classified as unaffected. Tests of association were performed using the Unphased program [147], which, in addition to family-based tests, is able to jointly examine pedigrees and case/control samples.

The analysis performed by the Unphased program is likelihood-based (section 2.4.3). The program defines separate association parameters in the parental and offspring components of the likelihood, allowing the introduction and conditioning on a so-called inheritance vector, which maintains robustness of the test to linkage when there are multiple affected offspring present in a nuclear family. The final statistic is calculated from a likelihood ratio test, which compares a model "with" against a model "without" the SNP of interest.

Unrelated subjects are regarded as the children of two missing parents and are then included in the same formulation as nuclear families [147].

Unphased accounts for missing data by use of a score function (section 2.4.3) of the parental genotypes model [147].

A Bonferroni correction was applied to account for multiple testing (section 2.4.4). Importantly, the association test results for SNPs genotyped in both the Cardiff University and Duke University cohorts are only reported for combined analyses. The implications of this approach, with respect to potential population stratification between subjects from the

UK and USA, are discussed in section 4.4.

4.3. Results

4.3.1 Subjects and Genotyping

The UK cohort comprised of 164 families with high myopia (604 subjects), along with an additional set of unrelated individuals comprised of 112 highly myopic cases and 114 "emmetropic" controls (spherical equivalent refractive error in both eyes > −1.00 D and < +1.00 D).

The USA cohort comprised of 86 families with high myopia (358 subjects), along with an additional set of 56 highly myopic unrelated individuals.

The combined study population included a total of 1251 subjects. Forty-nine subjects were excluded due to genotyping failure. This left 293 unrelated and 909 related individuals available for association analyses (Table 4.4): 788 subjects in the UK cohort (142 families, 121 cases and116 controls) and 414 subjects in USA cohort (86 families and 56 cases). Subjects for whom all relatives failed to pass our genotyping quality control threshold were classified as cases or controls if they met the necessary refractive criteria.

Genotyping of the two microsatellite markers NGA17 and NGA19 revealed four alleles for each. For both markers there were three common alleles and one rare allele. The observed allele frequencies of the microsatellite polymorphisms are shown in Table 4.3. Since the sample size was modest, the rare allele of each microsatellite marker was combined with the allele next in size to it (allele 1 with allele 2, for both markers).

Genotyping of SNPs had an average failure rate of ~7.5% (see Table 4.5 for individual rates per SNP). Concordance was assessed based on 8 duplicate samples: genotypes of 2 samples out of these 96 had discordant results. Genotyping for SNP marker rs235875 failed.

Table 4.4 Numbers of Subjects in the Study of Association Between High Myopia and Myocilin Gene

	Subjects (families) participating		Subjects (families) analyzed	
	UK	USA	UK	USA
Related	604 (164)	358 (86)	551 (142)	358 (86)
Cases	112	56	121	56
Controls	114	0	116	0
Total	830	414	788	414

Table 4.5 Tests of Association between MYOC Polymorphisms and High Myopia

Polymorphism	**Failed genotypes (%)**	**HWE p-value**	**Unphased p-value (corrected p-value)**	**Unphased Odds Ratio (95 % CI)**
Duke University Cohort				
rs6425363	1.5	1.00	0.57	1.15 (0.71-1.86)
rs235917	4.4	0.55	0.49	1.13 (0.79-1.59)
rs235875	2.7	0.20	0.36	1.20 (0.81-1.75)
rs2032555	3.5	0.01	*Not tested due to HWE status*	
Cardiff University Cohort				
rs235877	12.0	0.09	0.57	1.07 (0.84-1.37)
rs235870	9.0	0.27	0.53	0.93 (0.74-1.17)
rs2236875	10.0	0.01	*Not tested due to HWE status*	
rs235918	8.0	0.19	0.53	1.07 (0.86-1.34)
rs11586716	8.6	0.13	0.38	0.73 (0.84-1.44)
rs2075648	9.8	0.07	0.59	0.91 (0.64-1.28)
NGA17	0.1	0.08	0.03 (0.39)	0.70 (0.55-0.92)
NGA19	0.2	0.49	0.97	1.02 (0.82-1.26)
Combined Cohorts				
rs16864720	7.9	0.85	0.04 (0.645)	1.30 (1.004-1.73)
rs7545646	12.0	0.05	0.06	1.30 (0.98-1.8)
rs12076134	9.4	0.81	0.09	1.20 (0.97-1.48)
rs235858*	13.0	0.86	0.87	1.02 (0.84-1.22)
rs2421853*	13.0	0.18	0.25	1.13 (0.91-1.39)

* These two SNPs were significantly associated with high myopia in the study of Tang et al. [214].

4.3.2 Statistical Analysis

Tests for HWE showed that 2 SNPs, rs2236875 and rs2032555, were not in equilibrium in unrelated subjects (Table 4.5). Therefore, these two markers were dropped from further analyses, and association tests were performed for the remaining 15 variants: 13 SNPs and 2 microsatellites.

There was no significant heterogeneity in genotype frequencies between families and singleton samples either within or between cohorts (Table 4.3). Thus, families and unrelated subjects were analyzed jointly [147]. Likewise, subjects recruited at Duke University and Cardiff University were analyzed jointly for those SNPs genotyped in common (i.e. ignoring potential population stratification issues). The association test results are shown in Table 4.5. Prior to correction for multiple testing, two variants showed significant association: rs16864720 ($p=0.043$) and NGA17 ($p=0.026$). However, neither association retained statistical significance after Bonferroni correction (Table 4.5). Evaluation of relative risk highlighted the same two polymorphisms, rs16864720 and NGA17, with 95% confidence intervals that did not include 1.0 (Table 4.5). The relative risk conferred by each of these variants, however, was low ($RR < 1.5$).

4.4 Discussion

A joint analysis of subjects from the UK and USA was carried out for those SNPs that were genotyped in both groups of subjects. This pooling of subjects could potentially have given rise to a "false positive" or "false negative" association due to population stratification. However, population stratification can only give rise to a significant association between a disease phenotype and a marker genotype if (a) the prevalence of the disease differs between the two subject groups, *and* (b) the marker of interest's allele frequency differs between the two subject groups. For high myopia, exact figures on the prevalences in Caucasian subjects from the UK and USA are lacking, but estimates suggest these rates are similar [173, 341, 342]. Furthermore, the MYOC polymorphisms studied here had statistically similar allele frequencies in the UK and USA subjects (Table 4.3B).

In contrast to previously published significant associations between MYOC gene polymorphisms and high myopia [214, 319] in subjects of Chinese ethnicity, the present study suggested that there was no such relationship in subjects of Caucasian ethnicity. Indeed, the only polymorphisms that showed significant association before Bonferroni correction are situated at the 5' end (NGA17) and in the "middle" (rs16864720) of the gene rather than towards the 3' region implicated in the study of Tang et al. [214]. The ethnic difference of the respective study populations is an appealing explanation for these discrepant findings. Different populations may exhibit differences in allele frequencies and linkage disequilibrium patterns at specific loci (Figures 4.2, 4.3 and Table 4.3). Thus, the role of MYOC in high myopia in Chinese subjects may be dissimilar to that in Caucasians.

An alternative explanation could be the power of the analyses. The estimated relative risk of the genetic variants examined here was less than 1.5, which suggests that the power of this study would be ~75% [343]. Tang et al, on the other hand, investigated a smaller sample size (557 individuals, in 162 nuclear families) and reported a relative risk of >1.5 for two significant SNPs (rs235858 and rs2421853). To gain 80% power, a family based association study of a variant with relative risk > 1.5 and allele frequency of 0.5, would need ~200 families under an additive model and ~1100 families under a dominant model [344].

The fact that MYOC polymorphisms are implicated in both myopia and glaucoma is intriguing, especially in light of the higher-than-chance co-occurrence of myopia and glaucoma seen in many studies [320-322]. Nonetheless, the high expression of myocilin in the TMC [345] is easier to reconcile with the role of MYOC polymorphisms in glaucoma than in myopia. Furthermore, the current evidence suggests that the MYOC gene variants which confer an increased risk of open angle glaucoma are different from those that may increase susceptibility to myopia. In this respect, the association of MYOC polymorphisms with both conditions may be coincidental.

4.5 Conclusion

In conclusion, this study found no evidence to support a significant association between MYOC polymorphisms and high myopia in Caucasian subjects from the UK and USA.

CHAPTER V.

ASSOCIATION BETWEEN SNPS IN MYP REGIONS AND HIGH MYOPIA

5.1 Introduction

MYP regions are chromosomal intervals linked to myopia (section 1.4). Therefore, these genetic loci include or harbour a number of possible myopia genes. Several association studies - fine-mapping of linkage loci as well as candidate-gene analysis – have been carried out in attempt to identify genetic variants that confer susceptibility to myopia (section 1.4). To date, genes in the MYP 1, 2, 3, 5, 7 and 8 regions have been assessed. Some of the genes situated in these regions were only sequenced and screened for mutation, with no association test performed as such. The following genes have been investigated in detail: the TestisExpressed28 (TEX28) gene in MYP1 [195], the Lipin 2 gene [200, 202] and the leucin-rich repeat protein genes in the MYP2 [200, 201] and MYP 3 [201] regions. These studies found only suggestive [195, 202] or no evidence [200, 201] of a relationship with myopia.

Association tests have also been performed for intervals of MYP 2, 3, 5, 7 and 8, for genes such as Transforming Growth β-Induced Factor (TGIF), Lumican, Collagen type one alpha one (COL1A1), PairedBox6 (PAX6) and SexDeterminingRegionYBox 2 (SOX2). Apart from being situated within MYP loci, these genes also attracted myopia geneticists because of their biological role in embryonic development or, potentially, in myopic scleral remodelling. Although some of these studies have shown significant association, replication analyses have been disappointing (section 1.4), suggesting that the candidate gene(s) remain to be discovered for the most part.

In this chapter, exploratory association analyses are described for SNPs in all of the known MYP regions identified at the time of the study.

5.2 Materials and Methods

5.2.1 Subjects and DNA Samples

Recruitment of subjects and their DNA samples was accomplished within The International Myopia Consortium established in collaboration with myopia research centres in Denmark, Australia, France and USA [161]. DNA samples were collected in the form of mouthwash, blood and saliva. Objective or subjective measurements for spherical equivalent were obtained from each participating centre. The collection of high myopia pedigrees from Cardiff is described in sections 2.1 and 2.2.1.

5.2.2 Genotyping and Selection of SNPs

Whole genome genotyping was completed by the Center for Inherited Disease Research (CIDR; http://www.cidr.jhmi.edu/) using Illumina Linkage Panel IVb bead array system (http://www.illumina.com/pages.ilmn?ID=191). The average interpolated genetic map distance between all SNP loci was 0.62 cM.

For association analysis, SNPs were selected that are positioned within known genes in MYP regions or that are in high LD (r^2 >0.8) with such genes (Table 5.1). LD information and minor allele frequency for each SNP was obtained from the HapMap dataset for European (CEU) subjects. Those SNPs whose MAF in HapMap CEU subjects is less than 5% were excluded from the analysis.

5.2.3 Statistical Analyses

5.2.3.1 Hardy-Weinberg Equilibrium and Mendelian Consistency

The Pedstats package [340] was used to carry out an exact test for Hardy-Weinberg equilibrium (HWE) on unrelated subjects and to check for Mendelian consistency in pedigrees (section 2.3.3).

5.2.3.2 Association Tests (APL) and Correction for Multiple Testing

High myopia was examined as a dichotomous trait. Subjects with a spherical equivalent refractive error, averaged between eyes, of < -6.00 D were classified as affected [339]. All other subjects were classified as unaffected. Participants whose refractive error was not obtainable were coded as unknown.

Association tests were performed using the APL (Association in the Presence of Linkage) program, which calculates identity by descent (IBD – the probability that two individuals in a pedigree possess the same allele inferred from a recent common ancestor) parameters, and which can account for the presence of linkage when testing for association and/or inferring missing parental genotypes [149].

The basic concept of APL is that if there is no association, then the expected number of copies of alleles at the examined marker in siblings, given the genotypes of their parents, would be the number of copies of the marker alleles in the parents. APL employs the standard likelihood theory (section 2.4.3) with additional parameters of probabilities that the affected siblings share 0, 1 or 2 alleles IBD respectively at the marker locus [149]. In addition, it has been demonstrated that under the null hypothesis of no association and in the presence of missing parental data, APL showed greater power than other family-based association tests [346]. A Bonferroni correction was applied to account for multiple testing (section 2.4.4).

Table 5.1 Summary of SNPs Chosen for the Exploration of MYP Regions

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2–value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs1024694	MYP 1	FMR1NB	FMR1 (0.89)	0.489	None	0.8791	0.9931
rs1860929	MYP 1	AFF2	None	0.278	Corrected	0.5258	0.3144
rs758439	MYP 1	AFF2	None	0.278	Corrected	0.7135	0.1781
rs985595	MYP 1	AFF2	None	0.200	**Many**	Excluded	Excluded
rs222398	MYP 1	MTM1	None	0.389	None	0.7598	0.2336
rs6526192	MYP 1	MTMR1	HMGB3 (1.00)	0.167	Corrected	0.6688	0.6712
rs770238	MYP 2	LPIN2	None	0.325	Corrected	1.0000	0.8035
rs643015	MYP 2	LPIN2	None	0.308	Corrected	1.0000	0.9803
rs168206	MYP 2	DLGAP1	None	0.424	Corrected	0.8103	0.1006
rs1565728	MYP 3	E2F7	None	0.475	None	0.7074	0.6612
rs998070	MYP 3	NAV3	None	0.483	None	0.3995	0.1612
rs1351214	MYP 3	NAV3	None	0.417	None	0.6299	0.7881
rs2404772	MYP 3	None	CART1 (0.95) LRRIQ1 (1.00)	0.250	None	0.0871	0.6896
rs1508595	MYP 3	None	KITLG (1.00) None	0.167	None	0.1344	0.8446
rs1401982	MYP 3	ATP2B1	SYCP3 (1.00) GNPTAB	0.417	None	0.1306	0.1065
rs1544921	MYP 3	CHPT1	(1.00) FLJ11259 (0.88) None KIAA1033 (0.96)	0.475	None	0.6260	0.5363
rs746035	MYP 3	CHST11	None	0.308	None	0.2459	0.8233
rs9143	MYP 3	DIP13B	None	0.415	None	0.6300	0.7979
rs1922438	MYP 3	RFX4	None	0.475	None	0.1018	0.8294
rs2873108	MYP 4	DPP6	None	0.161	None	0.8668	0.9694
rs306278	MYP 4	DPP6	None	0.450	None	0.1118	0.3038
rs2033108	MYP 5	None	PCTP (1.00)	0.242	Corrected	0.7422	0.2009
rs759109	MYP 5	ANKFN1	None	0.425	Corrected	0.6281	0.6572
rs1024819	MYP 5	MSI2	None	0.492	Corrected	0.3361	0.9228
rs1974692	MYP 5	MSI2	None	0.195	Corrected	0.8511	0.3547
rs13137	MYP 5	None	TMEM49 (1.00)	0.212	Corrected	0.1076	0.3472
rs1881441	MYP 5	CLTC	None	0.117	Corrected	**0.0230**	Excluded
rs1557720	MYP 5	BRIP1	None	0.458	Corrected	0.2190	0.3326

Table 5.1 Summary of SNPs Chosen for the Exploration of MYP Regions (Continuation)

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs1997719	MYP 6	EMID1	None	0.367	Corrected	0.6249	0.8585
rs140062	MYP 6	EWSR1	None	0.317	Corrected	0.4567	0.0563
rs715494	MYP 6	AP1B1	EWSR1 (0.95) GAS2L1	0.367	Corrected	0.5449	0.2604
rs714027	MYP 6	HORMAD2	(0.95)	0.433	None	0.5439	0.5232
rs737805	MYP 6	None	None	0.358	Corrected	**0.0043**	Excluded
			TBC1D1OA (1.00) SF3A1 (1.00) LOC550631 (1.00) LOC200312 (0.90) SEC14L2				
rs4444	MYP 6	OSBP2	(0.96)	0.442	Corrected	0.4672	0.1321
rs136488	MYP 6	RFPL2	None	0.475	Corrected	0.9048	0.8280
rs762883	MYP 6	SYN3	SLC5A4 (0.96)	0.400	Corrected	0.3958	0.8603
rs9862	MYP 6	SYN3	None	0.467	Corrected	0.0537	0.9840
rs138777	MYP 6	TOM1	None	0.333	Corrected	0.6203	0.7873
rs739096	MYP 6	MYH9	HMG2L1 (1.00)	0.458	Corrected	0.2786	0.4960
rs933224	MYP 6	MYH9	None	0.358	Corrected	0.6809	0.3662
rs2413411	MYP 6	CACNG2	None	0.325	Corrected	0.6926	0.2305
rs760519	MYP 6	NCF4	None	0.258	Corrected	0.0665	0.6827
rs1534880	MYP 6	CSF2RB	FLJ90680 (0.94) None	0.492	Corrected	0.9035	0.4905
rs4348874	MYP 7	PTPNS	None	0.274	Corrected	0.7657	0.7621
rs730348	MYP 7	NAV2	None	0.408	Corrected	0.6237	0.2601
rs1470251	MYP 7	NAV2	None	0.183	Corrected	0.7583	0.6671
rs1374719	MYP 7	SLC17A6	None	0.306	Corrected	0.3385	0.9208
rs2928345	MYP 7	GAS2	None	0.192	Corrected	0.5741	0.4831
rs1491846	MYP 7	None	KIF18A	0.167	None	0.2289	0.7189
rs1032090	MYP 7	METT5D1	(0.81)	0.375	Corrected	0.4733	0.6738
rs1564745	MYP 7	METT5D1	None	0.375	Corrected	0.4023	0.3121
rs524373	MYP 7	None	None	0.317	None	0.3416	0.8939
rs2273544	MYP 7	PC11L1	KCNA4 (1.00) None	0.280	Corrected	0.1383	0.6214
rs373499	MYP 7	CSTF3	None	0.423	Corrected	1.0000	0.0607

Table 5.1. Summary of SNPs Chosen For the Exploration of MYP Regions (Continuation)

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs765695	MYP 8	None	C3ORF58 (1.00)	0.442	Corrected	0.3323	0.4568
rs723490	MYP 8	None	C3ORF58 (1.00)	0.460	None	0.3959	0.4818
rs1707465	MYP 8	PLOD2	None	0.397	Corrected	0.2822	0.2893
rs1027695	MYP 8	None	ZIC4 (0.93)	0.494	None	0.3872	0.7378
rs1450344	MYP 8	None	TSC22D2 (1.00)	0.392	Corrected	0.1012	0.5613
rs1920395	MYP 8	P2RY14	None	0.325	Corrected	0.2795	0.5471
rs3863100	MYP 8	MDS1	None	0.075	Corrected	1.0000	0.7210
rs755763	MYP 8	MBNL1	None	0.25	Corrected	1.0000	0.8707
rs701265	MYP 8	P2RY1	None	0.217	Corrected	0.2470	0.6686
rs9438	MYP 8	DHX36DEAH	None	0.350	Corrected	0.7140	0.6188
rs1025192	MYP 8	MME	None	0.492	Corrected	0.7180	0.9215
rs359573	MYP 8	None	PLCH1 (1.00)	0.307	Corrected	0.6025	0.8617
rs986963	MYP 8	KCNAB1	None	0.317	Corrected	0.2895	0.6517
rs1384542	MYP 8	FLJ16641	None	0.333	Corrected	0.2654	0.0737
rs1074864	MYP 8	VEPH1	None	0.492	Corrected	0.9044	**0.0274**
rs1515628	MYP 8	SCHIP1	None	0.475	Corrected	0.6304	0.6792
rs1599386	MYP 8	None	PPM1L (1.00) NMD3 (0.83) SLITRK3 (1.00)	0.467	**Many**	Excluded	Excluded
rs920417	MYP 8	None	GOLPH4 (1.00)	0.458	Corrected	0.7181	0.8865
rs953834	MYP 8	None	None	0.065	None	0.6954	0.7665
rs877439	MYP 8	GOLPH4	None	0.500	**Many**	Excluded	Excluded
rs905129	MYP 8	TNIK	None	0.442	Corrected	1.0000	0.7908
rs1285082	MYP 8	FNDC313	None	0.458	Corrected	0.0685	0.4897
rs623021	MYP 8	AADACL1	None	0.450	Corrected	0.1101	0.9169
rs649695	MYP 8	NLGN1	None	0.275	Corrected	0.8830	0.6408
rs2046718	MYP 8	NLGN1	None	0.308	Corrected	0.5261	0.2334
rs753293	MYP 8	NAALADL2	None	0.267	Corrected	0.6062	0.7077
rs1549114	MYP 8	NAALADL2	None	0.358	**Many**	Excluded	Excluded
rs1468924	MYP 8	KCNMB3	None	0.342	Corrected	0.8967	0.1137
rs2049769	MYP 8	PEX5L	None	0.200	Corrected	0.4109	0.4009
rs1973738	MYP 8	KLHL6	None	0.375	Corrected	1.000	0.7779
rs2054172	MYP 8	KLHL24	None	0.233	Corrected	0.3788	0.1480
rs1401999	MYP 8	ABCC5	None	0.408	Corrected	0.4633	0.8593
rs869417	MYP 8	ABCC5	None	0.408	Corrected	0.3921	0.6833
rs4432622	MYP 8	VPS8	None	0.356	Corrected	0.7906	0.9974
rs3332	MYP 8	VPS8	None	0.390	None	1.0000	0.8922
rs6769709	MYP 8	LOC285382	None	0.075	Corrected	0.5273	0.7566
rs1837882	MYP 8	LIPH	None	0.458	Corrected	0.7151	0.2946
rs6808013	MYP 8	DGKG	None	0.208	Corrected	0.5077	0.1911
r1561026	MYP 8		None	0.267	Corrected	**0.0297**	Excluded

Table 5.1. Summary of SNPs Chosen For the Exploration of MYP Regions (Continuation)

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs1039559	MYP 9	TMEM156	None	0.458	Corrected	0.6305	0.2245
rs974734	MYP 9	None	TMEM156 (1.00)	0.500	Corrected	0.5473	0.2270
rs1046655	MYP 9	None	KLHL5 (1.00) WDR19 (1.00) RFC1 (1.00)	0.433	Corrected	1.0000	0.4620
rs2035383	MYP 9	APBB2	None	0.492	Corrected	0.8112	0.8105
rs790142	MYP 9	APBB2	None	0.308	Corrected	0.1018	0.5961
rs1565114	MYP 9	ATP8A1	None	0.331	Corrected	1.0000	**0.0164 (2.296)**
rs1504491	MYP 9	None	GABRG1 (1.00)	0.500	Corrected	0.1034	0.4311
rs1866989	MYP 9	GABRB1	None	0.467	Corrected	**0.0096**	Excluded
rs225160	MYP 9	None	SPATA18 (1.00)	0.417	Corrected	0.1175	0.2173
rs751266	MYP 9	FIP1L1	SCFD2 (1.00) CLOCK	0.408	Corrected	0.3924	0.3961
rs2538	MYP 9	None	(1.00) None	0.300	Corrected	0.4295	0.2472
rs140643	MYP 9	AASDH2	IGFBP7	0.431	Corrected	0.4699	0.1705
rs899631	MYP 9	POLR2B	(1.00) None	0.400	Corrected	0.4549	0.8055
rs1456860	MYP 9	LPHN3	SRD5A2L2	0.308	Corrected	0.7818	0.9528
rs1879323	MYP 9	None	(1.00) None	0.433	Corrected	0.5262	0.8729
rs1483720	MYP 9	SRD5A2L2	CENPC1	0.433	Corrected	0.3992	0.0622
rs1899130	MYP 9	None	(1.00) SULT1B1	0.333	Corrected	0.4976	0.2119
rs1560605	MYP 9	None	(1.00) C4ORF7	0.142	Corrected	0.1365	0.0228 (3.192)
rs2063749	MYP 9	None	(1.00) CSN3 (1.00)	0.325	Corrected	0.8003	0.6123
rs9131	MYP 9	MTHFD2L	None	0.350	Corrected	0.4574	0.8693
rs717239	MYP 9	FLJ25770	None	0.408	Corrected	0.2292	0.7715
rs1511817	MYP 9	SHROOM3	None	0.276	Corrected	0.8900	0.0511
rs1566485	MYP 9	SNOT6L	None	0.425	Corrected	1.0000	0.2050
rs1426138	MYP 9	None	BMP2K (0.93)	0.175	None	0.2465	0.2975

Table 5.1 Summary of SNPs Chosen for the Exploration of MYP Regions (Continuation)

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs13429	MYP 10	None	C8ORF42 (1.00)	0.417	Corrected	0.4026	**0.0180 (2.520)**
rs935559	MYP 10	C8ORF42	None	0.127	Corrected	0.7915	0.4968
rs922798	MYP 10	CSMD1	None	0.433	Corrected	0.9013	0.8315
rs732299	MYP 11	None	RAP1GDS1 (1.00)	0.358	None	1.0000	0.3646
rs501110	MYP 11	RAP1GDS1	None	0.342	None	0.4506	**0.0454 (6.356)**
rs749407	MYP 11	MTTP	None	0.358	None	0.6958	0.1277
rs716556	MYP 11	DNAJB14	None	0.450	None	1.0000	**0.0249 (3.486)**
rs871061	MYP 11	BANK1	None	0.317	Corrected	0.4546	0.5292
rs230490	MYP 11	NFKB1	None	0.375	Corrected	0.2579	0.0985
rs747559	MYP 11	None	MANBA (1.00)	0.458	Corrected	0.4449	0.1244
rs228617	MYP 11	UBE2D3	None	0.450	Corrected	0.7193	**0.0151 (2.114)**
rs223383	MYP 11	LOC150159	None	0.483	None	0.7192	**0.0168 (2.352)**
rs223334	MYP 11	OC150159	None	0.483	None	0.8105	0.2866
rs995387	MYP 11	DC2	None	0.392	Corrected	1.0000	0.5572
rs1865845	MYP 11	AGXT2L1	None	0.314	Corrected	0.2327	0.1640
rs243985	MYP 11	ENPEP	None	0.425	Corrected	0.0633	0.5970
rs1354680	MYP 11	ANK 2	None	0.258	Corrected	0.8851	0.1733
rs967099	MYP 11	ANK2	None	0.350	Corrected	0.2051	0.1982
rs1380931	MYP 11	UGT8	None	0.442	Corrected	0.3360	0.0460 (6.440)
rs1459062	MYP 11	USP53	None	0.467	Corrected	0.8100	0.7152
rs537111	MYP 13	GUCY2F	None	0.467	None	0.5440	0.6793
rs697829	MYP 13	NXT2	None	0.378	Corrected	0.3341	0.4114
rs926412	MYP 13	PAK3	None	**0.000**	None	Excluded	Excluded
rs1016231	MYP 13	DCX	None	0.221	None	**0.0080**	Excluded
rs3027802	MYP 13	GLT28D1	None	0.189	Corrected	**0.0174**	Excluded
rs7049660	MYP 13	ZCCHC16	None	0.189	Corrected	1.0000	0.8756
rs2040497	MYP 13	ZCCHC16	None	0.200	Corrected	0.0806	0.2821
rs583430	MYP 13	None	LHFPL1 (1.00)	0.167	Corrected	**0.0255**	Excluded

Table 5.1 Summary of SNPs Chosen for the Exploration of MYP Regions (Continuation)

Chosen SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2–value)	MAF	Mendelian errors	HWE p-value	APL p-value (corrected p-value)
rs1577454	MYP 13	HTR2C	None	0.156	Corrected	0.5717	0.1441
rs7056311	MYP 13	LRCH2	None	0.211	Corrected	0.3285	0.5906
rs2231	MYP 13	None	LUZP4	0.178	Corrected	0.1171	0.9725
rs988457	MYP 13	None	(1.00)	0.478	Corrected	0.1196	0.5118
rs1716758	MYP 13	None	SLC6A14 (1.00)	0.211	Corrected	0.0898	0.4427
rs929590	MYP 13	None	WDR44 (0.93) IL13RA1 (0.90) DOCK11 (0.80)	0.133	Corrected	1.0000	0.3165
rs2227098	MYP 13	GRIA3	None	0.467	Corrected	0.8795	0.6008

5.3 Results

5.3.1 Subjects and Genotyping

In total 1462 subjects were recruited by the five myopia centres. As phenotypic data and trio parental information were not obtainable for some subjects, 786 participants were available for the association tests. After checking for Mendelian errors, a further 3 pedigrees (15 subjects) were excluded, leaving 771 subjects for the final analysis: 101 subjects from Denmark (22 pedigrees), 3 subjects from Australia (1 pedigree), 299 subjects from USA (58 pedigrees), 204 subjects from the UK (46 pedigrees) and 164 subjects from France (36 pedigrees).

Genotyping had a reproducibility rate of 99.99% and a failure rate of 0.20% (based on 81 blind duplicates and assessed in the total cohort of 1462 subjects).

5.3.2 Statistical Analyses

Out of 152 chosen SNPs, 12 were excluded from the analysis: 4 SNPs were not Mendelian consistent, 7 SNPs were out of HWE ($P < 0.05$) and 1 SNP had a MAF less than 5% (Table 4.5).

Prior to correction for multiple testing, nine variants showed significant association: rs1074864 in MYP 8; rs1565114, rs1560605 and rs1511817 in MYP 9; rs13429 in MYP 10; rs501110, rs716556, rs228617, rs223383 and rs1380931 in MYP 11. However, none of them retained its statistical significance after Bonferroni correction (Table 4.5).

5.4 Discussion

A combined analysis of subjects from the UK, USA, Denmark, France and Australia was carried out for SNPs in MYP regions. Since family-based tests are robust against potential population stratification, this joining of subjects could not have given rise to a false positive result. If, on the other hand, a SNP was associated with high myopia in some population, but not in others, this joint analysis could have diluted the association signal.

Previously, several attempts have been made to find a relationship between high myopia and genes in MYP intervals [163, 167-169, 177, 185, 186, 198, 199, 203-208]. Despite some encouraging, positive findings in MYPs 3 [203, 204], 5 [205], 6 [206, 208] and 8 [186], replication of most of these results has failed [163, 167, 168, 177, 185, 198, 199, 203, 205, 207, 347]. Likewise, the association tests described in this chapter, revealed no significant relationship with high myopia in any of the 12 MYP regions examined. Nonetheless, these analyses did not permit the evaluation of any of the genes tested in the previous studies mentioned above.

The present study was performed using APL [149], whereas the assessment of an association between high myopia and myocilin (described in the previous chapter) was carried out with Unphased [147]. APL has been shown to be the most powerful test for association for family-based analyses in the presence of linkage [346], hence this method was applied to test for a relationship between MYP regions and high myopia. Analyzing myocillin as a candidate gene for high myopia, however, involved not only families, but also cases and controls, making APL unsuitable as it cannot accommodate both types of subjects.

One likely explanation for the high rate of unsuccessful association studies in general, and the present study in particular, is a lack of power. Assuming myopia is a multifactorial disease, the polymorphisms responsible for it would likely be of low effect (genotype relative risk < 1.5), requiring a large sample size for their detection (section 1.3.4.3.1). Most of the studies performed to date (including this one) had a sample size of 1000 subjects or less (an exception is Andrew and colleagues study [186] with 1430 participants), while to achieve ~80% power at alpha 0.001 level, a relative risk of 1.3 and an allele frequency of 20%, one would need ~6000 people [122].

Another factor that could have led to the failure of the present study is poor genome coverage: the detection of a genetic variant with an effect on disease susceptibility will suffer if genetic markers with only low LD with the variant of interest are genotyped. In my analyses of MYP regions, the genotyped SNPs were widely spaced (~0.62cM), and as such poorly suitable for fine scale association analysis and candidate gene association

studies. From the International HapMap Project, it has been determined that the vast majority of SNPs with MAF of at least 5% can be reduced to ~550,000 LD bins for individuals of European and Asian ancestry. By genotyping at least one tagging SNP from each bin (~550,000 SNPs), ~80% of SNPs present at a frequency above 5% across the genome can be covered [348, 349]. Thus, the 140 SNPs examined here – although having a high prior probability of association due to their location within MYP regions – represent a massively limited panel of markers.

An important aspect of attempting to fine map or replicate different association results is the population on which the relationship was originally tested. Previous significant findings have been mostly reported in Asian populations [197, 199, 204, 206]. Therefore, the replication of such results in Caucasian population may fail due to the genetic discrepancies between different populations (section 1.3.3.2).

The specification of myopia as a phenotype varies among studies: some of those reporting significant finding have categorised myopia affectation as < - 10.00 D [199] or as < - 9.25 D [204]. Thus, analyses defining myopia as < - 6.00 D may find no association.

An alternative explanation for my negative findings could also be that common variants do not explain a substantial proportion of the phenotypic variation in refractive error. For example, there are 18 common variants that have been associated with type 2 diabetes, with MAFs ranging from 0.073 to 0.50 and relative risk ranging from 1.05 to 1.37. Together, however, these 18 polymorphisms explain less than 4% of the total liability of the trait [116]. Under the "common disease, rare variant" hypothesis, the large levels of heritability could reflect the aggregate effects of very many, very rare variants: each potentially of moderate effect but accounting for virtually none of the variation at the population level [350]. Such polymorphisms could be analyzed using different approaches that do not assume that common variant underlies a disease: if multiple rare disease variants exist within the same genomic region, then, instead of a standard association test, a linkage analysis-like approach of examining ancestral sharing at a locus performed in population-based samples of unrelated individuals may be considered [260].

It is also important to highlight, that in contrast to case/control studies, family-based tests of association are confounded with tests of linkage: marker/disease association will only be detected if the marker and causative allele are in strong LD and, in addition, are linked [351]. Moreover, in large genome-wide studies, the case/control design has been proved to be a more powerful tool of testing for association than a family-based method [352]. Therefore, this analysis of association between MYP regions and high myopia could have been more beneficial if performed in a case/control dataset rather than in families.

5.5 Conclusion

This study failed to find a significant relationship between high myopia and SNPs in 12 MYP intervals, most probably due to the lack of statistical power and poor genome coverage of the SNP panel.

CHAPTER VI.

ASSOCIATION BETWEEN HIGH MYOPIA AND COLLAGEN GENES

6.1 Introduction

The human eye contains a wide diversity of connective tissues, including the cornea, sclera, trabecular meshwork and vitreous, that function in a coordinated manner to ensure clear vision [353]. Genetic disorders involving connective tissues generally have a profound effect on the eye: such conditions as Marfan or Stickler syndromes have severe myopia as a consistent phenotype. It is, therefore, anticipated that abnormalities of the eye's connective tissue will result in impaired vision.

The connective tissues of the eye consist mostly of a collagenous extracellular matrix (ECM) network, the major part of which is composed of bundles of collagen fibres, surrounded by a complex matrix of proteoglycans and glycoproteins [354]. The lamellar collagen fibril bundlesare secreted and maintained by so-called fibroblast cells [354, 355]. Fibroblasts are thought to be able to differentiate to myofibroblasts through either stress or stimulation with signalling growth factors such as the cytokine transforming growth factor beta (TGFβ) [356]. Facilitated by direct cell-matrix interactions, myofibroblasts are able to modify their surrounding ECM both by contraction and the production of collagen, proteoglycans and many other constituents and regulatory molecules [357].

As myopia is mostly due to the elongation of vitreous chamber of the eye (section 1.2.1.2), in this chapter associations between variants in the major collagen types found in the sclera (COL1A1) and the vitreous (COL2A1) with high myopia were examined.

The relationship between the sclera and myopia is not clear. However, being significantly thinner in myopic than in non-myopic eyes, the sclera is considered to be an important component in myopic eye growth [355, 358-360]. Generally, loss of collagen and proteoglycans in ECM is thought to be responsible for myopic scleral thinning [358-360]. Studies examining scleral changes in myopia development highlight the altered expression of various fibrillar collagens (types I, III and V) [361-363]and the involvement of matrix-degrading enzymes (e.g. matrix metalloproteinases).

In sclera, collagen accounts for 80% of the dry weight, and is responsible for the strength and resilience of the tissue; 90% of the scleral collagen is of type I [364]. Animal studies have reported that mRNA expression of type I collagen in the sclera is reduced during

myopia development [361, 365]. Moreover, mutations in the collagen type-I gene (COL1A1) that encodes alpha one chains have been reported in clinical conditions associated with myopia, such as type-I osteogenesis imperfecta and the Ehlers-Danlos syndrome [366]. COL1A1 is located on chromosome 17 and consists of 51 exons (Figure 6.1). Its position within the MYP5 region further suggests a possible link with high myopia. In addition, significant association between two COL1A1 polymorphisms and myopia was found in a Japanese case/control study [205]. However, this initial positive finding failed to replicate in a different Japanese cohort and in a Chinese cohort [169, 347].

The vitreous body plays an important role in emmetropization and, thus, in the development of myopia. There is a rapid increase of the length of the vitreous chamber during the first year of life up to the age of 3 years, followed by a comparatively slow increase from the age of 3 to 12 years [367].

The vitreous body consists of the vitreous cortex and the central vitreous. The vitreous cortex is a thin layer of dense collagen fibrils, running parallel to the retina and attached to its internal membrane. The major constituent of the central vitreous (also known as vitreous humor) is water, although it exists in the form of gel. The pivotal role in maintaining this gel structure is played by a low concentration of collagen fibrils [368]. It has been proven that removal of these collagen fibrils results in the conversion of the gel into a viscous liquid [367]. Liquefaction of vitreous gel, however, occurs physiologically during aging [369]. In patients with high myopia this process begins at younger age than in non-myopic eyes and progresses with axial elongation, thus, resulting in a frequent occurrence of posterior vitreous detachment and, consequently, can lead to blindness [370].

Collagen fibrils of the vitreous are mostly of type II [353] and polymorphisms in the collagen type two alpha one (COL2A1) gene have been found to be associated with low grade myopia in a Caucasian population [163]. In addition, COL2A1 mutations are associated with Stickler syndrome, a condition that has severe myopia as a consistent phenotype [371]. The COL2A1 gene is located on chromosome 12 and consists of 54

exons (Figure 6.2).

In this chapter, replication of the previous significant association between COL1A1 and COL2A1 and high myopia was attempted in families and cases/controls recruited as part of the Family Study of Myopia. Analyses were performed by our group as well as by the research team of Prof. T.L. Young in Duke University (USA).

6.2 Materials and Methods

6.2.1 Subjects and DNA Samples

Subjects were recruited and DNA samples collected as described in sections 2.1 and 2.2.1. Apart from pedigrees with high myopia, unrelated cases and controls were also collected for these analyses. Whilst the ascertainment of pedigrees was carried out via optometrists/ophthalmologists throughout UK (section 2.1), cases and controls were recruited from the Cardiff University Eye Clinic's database only.

Individuals with known syndromic disorders or a systemic condition that could predispose to myopia were excluded. All subjects were of Caucasian ethnicity (self-reported "White Europeans").

6.2.2 Selection and Genotyping of SNPs

Being a SNP replication study, the SNPs examined were those found to be associated with myopia in previous studies: rs2075555 and rs2269336 in COL1A1 [205]; rs1635529 in COL2A1 [163] and rs1034762 and rs1793933 in COL2A1 by the group of Prof.T.L.Young in Duke University (USA).

Diagrams of positions of the selected SNPs in COL1A1 and COL2A1 genes are shown in figures 6.1 and 6.2. SNP genotyping was carried out by Kbiosciences Ltd.

6.2.3 Statistical Analyses

6.2.3.1 Hardy-Weinberg Equilibrium and Mendelian Consistency

The Pedstats package [340] was used to carry out an exact test for Hardy-Weinberg equilibrium (HWE) on unrelated subjects and to check for Mendelian consistency in pedigrees (section 2.3.3).

6.2.3.2 Association Analysis and Correction for Multiple Testing

Association analyses were carried out by two researchers: one from the group of Prof. Terri L. Young in Duke University (USA) and myself at Cardiff University. Both teams analyzed the same cohort collected as a part of The Family Study of Myopia. Analyses completed in Duke University did not involve me.

Duke University Analysis: High myopia was examined as a dichotomous trait. Subjects with a spherical equivalent refractive error, averaged between eyes, of < -5.00 D were classified as affected [172]. All other subjects were classified as unaffected, while subjects whose refractive error was unobtainable were coded as unknown. The researcher in this group chose to perform the genetic analysis using PDT (Pedigree Disequilibrium Test [372]) and to analyze pedigrees only to avoid issues related to population stratification.

Cardiff University Analysis: High myopia was examined as a dichotomous trait. Subjects with a spherical equivalent refractive error, averaged between eyes, of < -6.00 D were classified as affected [339]. All other subjects were classified as unaffected, while subjects whose refractive error was unobtainable were coded as unknown. Due to the expected modest size of samples that comprises pedigrees only, I carried out association tests on a combined cohort of families and cases/controls using the Unphased genetic program (section 4.2.3.2).

A Bonferroni correction was applied to correct for multiple testing in the analyses of both groups. The correction was performed for 10 tests: 5 SNPs examined in families only and an additional 5 tests for the same SNPs performed in the combined dataset.

Figure 6.1 COL1A1 gene position, structure, LD patterns and genotyped SNPs

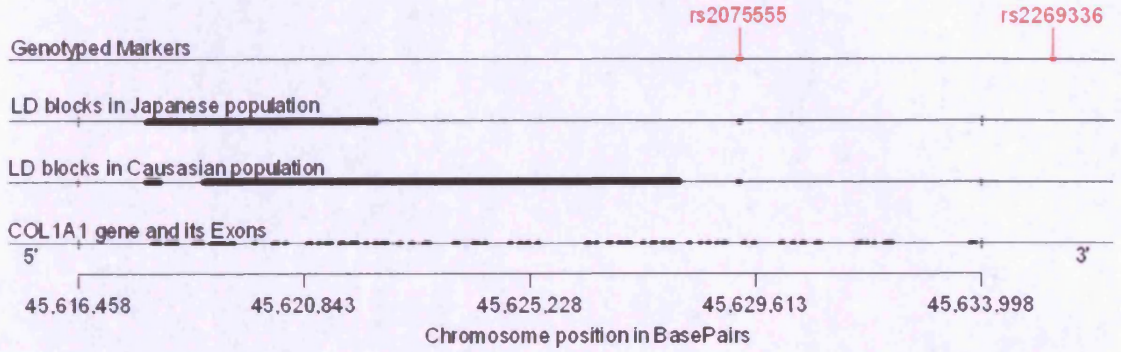
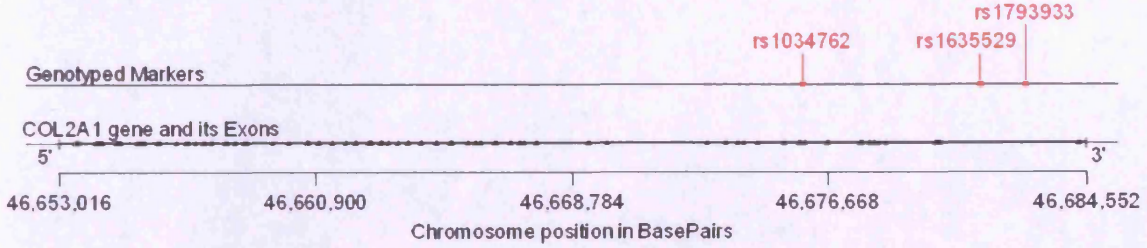


Figure 6.2 COL2A1 gene position, structure and genotyped SNPs



6.3 Results

6.3.1. Subjects and Genotyping

Genotyping of SNPs had an average failure rate of ~7.5 %. Genotyping concordance was assessed to be 99.38% based on 8 duplicate samples. Allele frequencies of the genotyped SNPs are shown in Table 6.1.

The Duke University researchers analyzed 130 multiplex families (582 individuals), while our group examined 164 families with high myopia (604 subjects) and an additional set of unrelated individuals, comprising of 112 highly myopic cases and 114 "emmetropic" controls (spherical equivalent refractive error in both eyes > -1.00 D and $< +1.00$ D). (Note that the multiplex pedigrees analysed by both groups were largely identical – the difference in the number of pedigrees/subjects examined differed only because of me choosing to include cases/controls as well as some additionally recruited pedigrees into my analysis.)

6.3.2 Association Analysis

All SNPs were Mendelian consistent and in HWE (Table 6.2).

Duke University Analysis: The PDT analysis performed in families-only by the Duke team revealed a significant association between rs1635529 in COL2A1 and high myopia (Table 6.2). The initial p-value of 0.0044 remained statistically significant at the 0.05 level after Bonferroni correction (p=0.044). There was no significant association between any of the COL1A1 SNPs and high myopia.

Cardiff University Analysis: In contrast to the significant PDT result of Prof. T.L. Young's group, my analysis of the joint family-plus-case/control sample, using the Unphased program, failed to show an association with high myopia and any SNP in either COL1A1 or COL2A1 (Table 6.3).

Table 6.1 Allele frequencies of the genotyped SNPs in COL1A1 and COL2A1

Gene	SNP	Allele	Frequency among founders in families	Frequency among cases	Frequency among controls
COL1A1	rs2075555	T	0.1321	0.1351	0.1538
		G	0.8679	0.8649	0.8462
	rs2269336	C	0.1215	0.1327	0.1415
		G	0.8785	0.8673	0.8585
COL2A1	rs1635529	T	0.1611	0.1589	0.1727
		G	0.8389	0.8411	0.8273
	rs1034762	A	0.1722	0.1500	0.1698
		C	0.8278	0.8500	0.8302
	rs1793933	G	0.1770	0.1606	0.1847
		T	0.8230	0.8394	0.8153

Table 6.2 Results of the Replication Study between High Myopia and COL1A1 and COL2A1 polymorphisms performed by the group of Prof. T.L.Young in Duke University (USA)

Gene	SNP	HWE p-value	PDT p-value (corrected p-value)
COL1A1	rs2075555	0.4053	p>0.05
	rs2269336	0.3916	p>0.05
COL2A1	rs1635529	0.4811	**0.0044 (0.044)**
	rs1034762	0.7247	0.0629
	rs1793933	0.4042	0.1175

Table 6.3 Results of the Replication Study between High Myopia and COL1A1 and COL2A1 polymorphisms performed by myself. All p-values are uncorrected.

Gene	SNP	HWE p-value	Unphased p-value	Odds Ratio	95% Confidence Interval
COL1A1	rs2075555	0.4053	0.4067	0.88	0.66 – 1.18
	rs2269336	0.3916	0.1022	1.29	0.95 – 1.75
COL2A1	rs1635529	0.4811	0.0640	0.76	0.56 – 1.02
	rs1034762	0.7247	0.2043	1.20	0.90 – 1.60
	rs1793933	0.4042	0.1521	1.22	0.93 – 1.61

6.4 Discussion

The replication of previously established significant association of myopia with polymorphisms in the collagen genes COL1A1 and COL2A1 was attempted in a Caucasian cohort. The association was examined in families only (performed by researchers in the Human Genetic Centre at Duke University, USA), and in a combined sample of families and unrelated cases/controls (performed by myself). Although the family-based analysis revealed a significant result for COL2A1, this was not confirmed in the combined dataset.

A likely explanation for the discrepant results is the different methods applied in these two analyses. The Duke University group examined families only and used the PDT program, while I analyzed families jointly with an additional case/control set of subjects and used the Unphased program for the calculation of association statistics. Although it has been shown that PDT (Pedigree Disequilibrium Test) has approximately the same power as Unphased under the null hypothesis of no association and in the presence of missing parental data [346], for a rare disease such as high myopia (prevalence ~ 2% in the European population), a trio-based design is sometimes more powerful than a case/control design [139], suggesting that examining families alone may be beneficial. In addition, I have used a different criterion for affectation status (more or equal to -6.0 Dioptres) compared to that of the team of Prof. T. Young (more or equal to -5.0 Dioptres). Because the analyses performed by me and by the Duke University group diverge at more than one point (number and type of participants, statistical software, affectation status threshold), it is difficult to pinpoint with certainty which one may be the reason for the observed opposition in these results.

The lack of consistency in the results of PDT and Unphased, suggests the possibility that the apparent COL2A1 replication, observed by the team of Duke University, may be a false positive result. According to the results I observed, the relative risk of the variant found significant by Duke group (rs1635529) is 0.76 (Table 6.3) and its minor allele frequency among founders is 0.16 (Table 6.1), suggesting that at a 0.05 significance level, the respective analyses had ~67% power when performed by the Prof.T.L.Young's team

and ~75% power when performed by myself [373]. Thus, a more powerful, larger sample size would be needed in order to confidently recognize a true association, if present.

Mutti and co-workers [163] established a significant association between rs1635529 in COL2A1 and myopia. Their study comprised of 123 families (517 individuals) and included subjects of varying ethnicity: Caucasian (62%), East Asian (13%), Hispanic (8%), African-American (7%), Indian/Pakistani (4%) and mixed or other ethnicity (6%). The group does not report an effect size for rs1635529, but calculates the minor allele frequency among the founders of examined families to be 0.21. Assuming the same effect as estimated in our study, the power of Mutti et al.'s analyses would be ~65% [373]. Although none of the analyses (performed on the families of The Family Study of Myopia or on the families collected by Mutti et al) have power of ~80% or more, the power of the test carried out by our group is 10% greater than that of Mutti et al's.

It is also important to point out that Mutti et al. established a positive relationship between COL2A1 and short-sightedness using an analysis model with myopia classified as < -0.75 D, suggesting that COL2A1 is linked to common myopia rather than to high myopia. The present study defined myopia as < -5.00 D (Prof.T.L.Young's group) or as < -6.00D (our group) and, thus, concentrated on a different phenotype. Therefore, this would not constitute true replication in a strict sense, because the phenotypes concerned were not identical.

Considering the arguments above, it is presumable that COL2A1 is associated with lower degrees of myopia; and/or the study of Mutti et al [163] suffers from a small power relative to that carried out here. In addition, because of the negative replication in the joint family and case/control analysis performed by our group, it is difficult to state conclusively that this replication was successful.

Another collagen gene examined in this study was COL1A1. Variants genotyped in this gene showed no significant association with high myopia independently of whether the team of Prof. Young or I performed the test (Tables 6.2 and 6.3). One of the explanations for this negative result could have been an ethnicity difference: the original report of

significant association between two COL1A1 SNPs and high myopia was found in the Japanese population [205], while we examined subjects of Caucasian background. SNPs which would be in strong LD with a disease-causing polymorphism in the COL1A1 gene in Japanese subjects would not necessarily be in strong LD with the disease-causing polymorphism in Caucasian subjects, due to the differing LD structure of COL1A1 in the two races (Figure 6.1). This racial difference in LD structure may also have been responsible for the significantly different minor allele frequencies (in Japanese versus Caucasian subjects) of the two SNPs that were originally implicated as disease causing ($p=0.004$ for rs2075555 and $p=0.0004$ for rs2269336; Fisher's test) as shown in Table 6.4. Thus, it was not entirely surprising that assessing the same genetic variants in these two different ethnic groups led to different results. Instead of a strict SNP replication study, a more thorough analysis of the COL1A1 gene, such as with tag SNPs chosen based on the LD pattern of the Caucasian population, would have been an option, which would have overcome this potential problem. However, the large size of the COL1A1 gene precluded this, as it would have been prohibitively expensive.

Comparing the genotype relative risks of the two significant SNPs (rs2075555 and rs2269336) of the Japanese study [205] and this one, rs2075555 has a slightly smaller effect (1.14 versus 1.30 in Japanese) in the results of our group, while rs2269336 has approximately the same odds ratio of 1.30 in both studies. The Japanese analysis involved 660 subjects in total (330 cases and 330 controls), while the test performed by our team was carried out on 830 participants (604 individuals in families, 112 cases and 114 controls). Given, that the relative risk of two examined SNPs is similar in two populations, it may be that the Japanese significant result established with a smaller sample size does not reflect a true association, but a type I error. This assumption is supported by two other, independent studies in the literature [169, 347] that also failed to replicate the original finding of association between COL1A1 and high myopia. Moreover, both of these negative replications were carried out on Asian population: one of them [347] involved Japanese participants (847 unrelated individuals) like the initial, positive test of Inamori et al [205]; and the other comprised of 1094 Han Chinese cases and controls [169].

Table 6.4 Allele Frequencies of COL1A1 polymorphisms in Japanese (JPT) and Caucasian (CEU) populations of HapMap

SNP	Allele	Frequency	
		JPT	CEU
rs2075555	T	0.367	0.183
	G	0.633	0.817
rs2269336	G	0.393	0.164
	C	0.607	0.836

Despite substantial research, contradictory results in genetic association analyses of candidate genes for such common, complex diseases as myopia are not rare. Explanations for this are the difficulty in recruiting very large sample sizes and poor genomic coverage. With modern whole-genome association studies examining a large number of SNPs, the later disadvantage can be overcome. However, from a set of hundreds of thousands of tests, many highly significant results are expected by chance alone, making it hard to distinguish a true signal from noise. This latter problem can only be solved by increases in sample size: fortunately, maintaining the same power when performing an exponentially larger number of Bonferroni-corrected tests requires only a linear increase in sample size. For example, if 500 individuals are needed to test a single SNP with an adequate power, then ~2,000 subjects would be suitable for testing 500,000 SNPs even after Bonferroni correction [260].

6.5 Conclusion

This study revealed a suggestive relationship between COL2A1 and high myopia. However, further, statistically more powerful analyses are needed to confirm this finding as a true positive association.

CHAPTER VII.

A TEST OF IMPRINTING IN HIGHLY MYOPIC CASE-PARENT TRIOS

7.1. Introduction

A gene is imprinted when its level of expression is dependent on the sex of the parent from whom it was inherited (section 1.3.1.3). Such gene expression contributes to resemblance between siblings as well as between parents and offspring [374], introducing a sex-specific element to the genetic mechanism of complex disorders.

The correlation of refractive error between relatives in families with high myopia has been well established [323, 375-377]. The between-sibling correlation has been estimated to vary from 0.31 [375] to 0.77 [376] in different populations. In addition, sister-sister correlation proved to be stronger than brother-brother or brother-sister correlations [323, 377], suggesting a potential sex (parent-of-origin) effect on refractive error. This effect is also supported by the observation that a female child tends to mirror the refractive error of her mother, while a male child mirrors the refractive error of his father [378].

In this study the possible effect of imprinting was examined in highly myopic offspring and their parents.

7.2 Materials and Methods

7.2.1 Subjects and DNA Samples

Trios (an offspring and its two parents) were selected from the families collected within The International Myopia Consortium (section 5.2.1). Complex families were reduced to trios in two steps: (1) first smaller, nuclear families with both parents genotyped were chosen; and then (2) one of the affected offspring from each family was randomly selected to form a trio. Only trios that had no missing data (both parents were available for recruitment) and had a highly myopic offspring were included in the analyses.

7.2.2 SNP selection and Genotyping

As described previously (section 5.2.2), a panel of SNPs was genotyped using the Illumina Linkage Panel IVb bead array system (http://www.illumina.com/ pages.ilmn?ID=191),

completed by the Center for Inherited Disease Research (CIDR; http://www.cidr.jhmi.edu/). From the approximately 6000 SNPs genotyped by CIDR, SNPs for the present analysis were selected in two stages: firstly, SNPs related to genes within established MYP regions were short listed (as described in section 5.2.2 and shown in Table 5.1); and secondly, SNPs with any missing genotype calls were excluded (this was done because the software used to perform the test of imprinting effect can not handle any missing data).

7.2.3 Statistical Analysis

7.2.3.1 Hardy-Weinberg Equilibrium and Mendelian Consistency

The Pedstats package [340] was used to carry out an exact test for Hardy-Weinberg equilibrium (HWE) on unrelated subjects and to check for Mendelian consistency in pedigrees (section 2.3.3).

7.2.3.2 Test of Imprinting (TRIMM) and Correction for Multiple Testing

High myopia was examined as a dichotomous trait. Subjects with a spherical equivalent refractive error, averaged between eyes, of < -6.00 D were classified as affected [339]. All other subjects were classified as unaffected, while subjects whose refractive error was unobtainable were coded as unknown.

The analyses were performed with the TRIMM package. This software was developed for testing for parent-of-origin effects in case-parent trios [379]. TRIMM first examines the transmission of alleles from parents to offspring by constructing, for each trio, a so-called "complementary sibling" who carries the two alleles not transmitted to the (real) affected child. It then computes a "difference vector" between the alleles transmitted to the affected offspring cases and to the complementary siblings. Under the null hypothesis that the set of markers is not associated with disease status within families, the genotype distributions of cases and their complements are the same, and, consequently, the expected value of the difference vector is zero. (Thus, this first step assesses transmission distortion,

reflecting the intuition that any set of alleles jointly related to risk will have been transmitted to the affected offspring more often than to the complement). The test computes a z-statistic (the value of a vector divided by its standard error) for each SNP in turn and identifies the maximum z-score ("Z_max") across all examined loci. The statistical significance is assessed using the permutation distribution of squared Z_max over random re-assignments of the labels "case" and "complement". In addition to z-statistics, a so-called T^2 – statistic is also calculated. The latter is used to exploit the correlation structure produced by LD and is expected to be beneficial when the causative SNP is not genotyped or, alternatively, the increased susceptibility is due to a particular set of SNPs. Statistical significance is assessed using the permutation distribution of T^2.

As a result of the first step, TRIMM produces two p-values for the possible transmission disequilibrium (Z_max and T^2 – statistic) and, thereby, identifies a set of SNPs transmitted from parents to offspring that potentially could play a role in susceptibility to the examined phenotype. Under the default settings, TRIMM assumes that SNPs are "potentially" over- or under-transmitted if the Z_max p-value is less than 0.1.

For its second step, TRIMM examines if there is a parent-of-origin effect. This is achieved by calculating another difference vector: in this case, the "SNP-count difference" between mothers and fathers. Assuming that one of the parental (maternal or paternal) alleles alone conferred the risk, then one of the parents would be more likely than the other to carry that risk allele, producing an asymmetry in maternal and paternal SNP-allele counts. Similarly, Z-statistic calculation and permutation testing is applied to nominate a parental set of risk SNPs for effects mediated through each parent. Before performing these calculations, however, the trios are stratified into two groups: one group consisting of trios whose offspring possess the set of risk alleles identified in TRIMM's first step, and another group who do not. The reason for this is that, if imprinting were present, the parents of possible carriers should show a SNP-allele count difference for that set of risk alleles, whereas parents of the definite non-carriers should not [379].

Figure 7.1 represents the two steps of TRIMM's algorithm.

Figure 7.1 Schematic Representation of TRIMM's Algorithm

Consider biallelic SNP with alleles 1 and 2. The four trios represent some possible combinations of genotypes for such SNP.

Stage One Analysis: Transmission distortion test performed on the offspring, resulting in the identification of the set of "offspring risk alleles". The difference vector D for this stage is calculated as 2C-(M+F), where C, M and F are the number of copies of designated allele at the SNP examined. This figure shows D vectors for each trio assuming allele 1 is the analyzed one. Z-statistic is determined as $D_{average}/SE$, where $D_{average}$ is the average D among informative families (the D vector is not zero) and SE is the corresponding standard error.

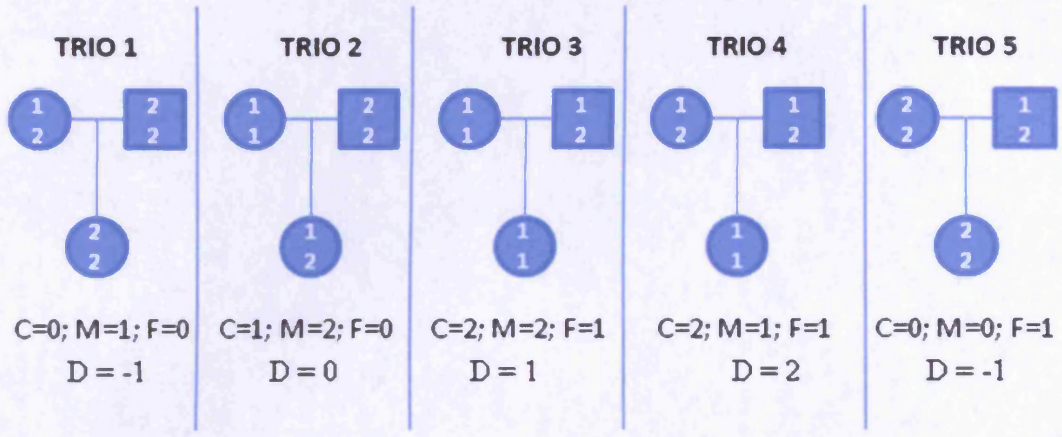
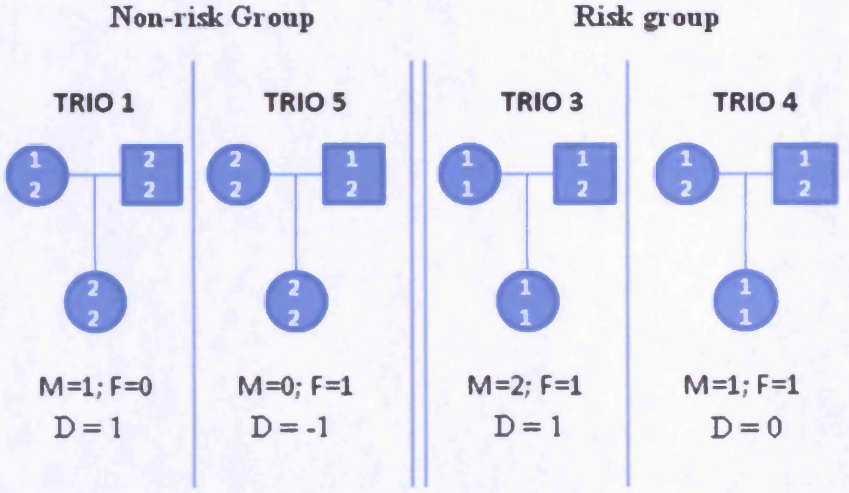


Figure 7.1 Schematic Representation of TRIMM's Algorithm (Continuation)

Stage Two Analysis: Test for SNP-allele counts in parents, resulting in the nomination of the set of "parental risk alleles". The difference vector D for this stage is calculated as M-F, where M and F are the same counts as in the stage one analysis. Trios are stratified according to whether the child possesses at least one copy of each nominated "offspring risk alleles" or not, creating a risk group and a non-risk group. Assuming that allele 1 of the SNP represented in this figure is nominated as the "offspring risk allele", risk group would consist of trios 2 and 4, while non-risk group would be trios 1 and 5. If the parental test of this stage is significant in the risk group only, the data supports a parent-of-origin effect. If, however, the test is significant in both groups, the data supports the presence of a maternal effect rather than a parent-of-origin effect.



7.3 Results

7.3.1 Subjects and Genotyping

The initial collection of pedigrees ascertained by the International Myopia Genetics Consortium comprised 1462 subjects. After the exclusion of subjects for whom phenotypic and/or parental data was not available, and of pedigrees that were not Mendelian consistent, there were 771 individuals in 264 pedigrees (section 5.3.1). An additional 498 participants were excluded as a result of (1) pedigrees being trimmed to trios; (2) trios not having an affected child; and (3) both parents not having been genotyped. Finally, 91 trios (273 subjects) were included in the dataset selected for TRIMM analysis.

As described in Chapter 5 (section 5.3.2), of the ~6000 SNPs genotyped in the Illumina panel, 140 SNPs were information with respect to genes within myopia MYP regions, passed the HWE quality control test ($p>0.05$) and were Mendelian consistent. Because TRIMM handles X-chromosome genotypes by stratifying the trios according to the sex of the offspring, SNPs on the X-chromosome were also excluded (because the small sample size would not allow enough power for such an analysis). In addition, as the T^2-statistic is not able to handle any missing data, SNPs with any failed genotypes were also excluded. This left 86 SNPs available for the TRIMM analyses.

7.3.2 Statistical Analyses

The first step of TRIMM showed little/borderline evidence of transmission distortion among affected offspring according to the Z_max statistic ($p=0.053$), and proved to be even less convincing after performing the T^2 – test ($p=0.163$). Nonetheless, individual tests of each SNP nominated a set of 6 "offspring risk SNPs" that potentially could confer an increased risk of high myopia (Table 7.1).

Trios were stratified according to whether the offspring carried at least one copy of the potential susceptibility allele of each SNP in the "offspring risk set". This resulted in 20 trios being assigned to the "risk group" and the remaining 71 trios being assigned to the

"non-risk group". After performing the second TRIMM analysis step – the test for a difference in SNP-allele count amongst parents – the z-score statistic showed no evidence of asymmetry in the "risk group" (p=0.256) but did suggest a significant result in the "non-risk group" (p=0.005). The T^2 – test, on the other hand, revealed a significant difference in both groups (p=0.035 in the risk group and p=0.012 in the non-risk group). TRIMM did not identify a "parental set of risk SNPs" for the risk group, since the overall z-statistic p-value did not reach the program's threshold p-value of 0.1 (Table 7.2). In the non-risk group, however, TRIMM identified a set of 4 "parental risk SNPs" (Table 7.3). The 6 "offspring risk SNPs" nominated in step one and the 4 "parental risk SNPs" nominated in step two consisted of different variants (i.e. there was no overlap in the list of risk SNPs identified).

Table 7.1 Test of Imprinting Results (Step One): Transmission Distortion to Affected Offspring

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	p-value (z-statistic)
rs770238	MYP 2	LPIN2	None	0.325	0.920
rs168206	MYP 2	DLGAP1	None	0.424	0.100
rs1565728	MYP 3	E2F7	None	0.475	1.000
rs998070	MYP 3	NAV3	None	0.483	1.000
rs2404772	MYP 3	None	CART1 (0.95) LRRIQ1 (1.00)	0.250	0.110
rs1508595	MYP 3	None	KITLG (1.00)	0.167	0.660
rs1401982	MYP 3	ATP2B1	None	0.417	0.450
rs1544921	MYP 3	CHPT1	SYCP3 (1.00) GNPTAB (1.00) FLJ11259 (0.88)	0.475	0.162
rs1922438	**MYP 3**	**RFX4**	**None**	**0.475**	**0.095 (1.793)**
rs2873108	MYP 4	DPP6	None	0.161	0.268
rs306278	MYP 4	DPP6	None	0.450	0.326
rs2033108	MYP 5	None	PCTP (1.00)	0.242	0.505
rs1024819	MYP 5	MSI2	None	0.492	0.527
rs1974692	MYP 5	MSI2	None	0.195	0.303
rs1881441	MYP 5	CLTC	None	0.117	1.000
rs1557720	MYP 5	BRIP1	None	0.458	0.466
rs715494	MYP 6	AP1B1	EWSR1 (0.95) GAS2L1 (0.95)	0.367	0.934
rs714027	MYP 6	HORMAD2	None	0.433	0.661
rs4444	**MYP 6**	**OSBP2**	**None**	**0.442**	**0.085 (-1.906)**
rs762883	MYP 6	SYN3	None	0.400	0.578
rs9862	MYP 6	SYN3	None	0.467	0.380
rs739096	MYP 6	MYH9	None	0.458	0.671
rs2413411	MYP 6	CACNG2	None	0.325	0.705
rs760519	MYP 6	NCF4	FLJ90680 (0.94)	0.258	0.597
rs1534880	MYP 6	CSF2RB	None	0.492	0.157
rs4348874	MYP 7	PTPNS	None	0.274	1.000
rs730348	**MYP 7**	**NAV2**	**None**	**0.408**	**0.002 (3.764)**
rs1470251	MYP 7	NAV2	None	0.183	0.410
rs1374719	MYP 7	SLC17A6	None	0.306	0.230
rs2928345	MYP 7	GAS2	None	0.192	0.830
rs1491846	MYP 7	None	KIF18A (0.81)	0.167	0.160
rs1032090	MYP 7	METT5D1	None	0.375	0.740
rs1564745	MYP 7	METT5D1	None	0.375	0.670
rs524373	MYP 7	None	KCNA4 (1.00)	0.317	0.660
rs2273544	MYP 7	TPC11L1	None	0.280	0.610
rs373499	MYP 7	CSTF3	None	0.423	0.520

Table 7.1 Test of Imprinting Results (Step One): Transmission Distortion to Affected Offspring (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	p-value (z-statistic)
rs765695	MYP 8	None	C3ORF58 (1.00)	0.442	0.140
rs723490	MYP 8	None	C3ORF58 (1.00)	0.460	0.140
rs1027695	MYP 8	None	ZIC4 (0.93)	0.494	0.250
rs1920395	MYP 8	P2RY14	None	0.325	0.570
rs755763	MYP 8	MBNL1	None	0.25	0.710
rs701265	MYP 8	P2RY1	None	0.217	0.600
rs9438	MYP 8	DHX36DEAH	None	0.350	0.900
rs1025192	MYP 8	MME	None	0.492	0.440
rs359573	MYP 8	None	PLCH1 (1.00)	0.307	0.560
rs986963	MYP 8	KCNAB1	None	0.317	0.210
rs1384542	MYP 8	FLJ16641	None	0.333	0.930
rs920417	MYP 8	None	SLITRK3 (1.00)	0.458	0.707
rs953834	MYP 8	None	GOLPH4 (1.00)	0.065	0.684
rs3863100	MYP 8	MDS1	None	0.075	0.869
rs905129	MYP 8	TNIK	None	0.442	0.846
rs1285082	MYP 8	FNDC313	None	0.458	0.815
rs2046718	MYP 8	NLGN1	None	0.308	0.910
rs753293	**MYP 8**	**NAALADL2**	**None**	**0.267**	**0.094 (-1.879)**
rs1468924	MYP 8	KCNMB3	None	0.342	0.241
rs2049769	MYP 8	PEX5L	None	0.200	0.727
rs1973738	MYP 8	KLHL6	None	0.375	1.000
rs1401999	MYP 8	ABCC5	None	0.408	1.000
rs869417	MYP 8	ABCC5	None	0.408	0.895
rs4432622	MYP 8	VPS8	None	0.356	0.324
rs3332	MYP 8	VPS8	None	0.390	0.288
rs1837882	**MYP 8**	**LIPH**	**None**	**0.458**	**0.026 (2.372)**
rs6808013	MYP 8	DGKG	None	0.208	0.471
rs1039559	MYP 9	TMEM156	None	0.458	0.177
rs974734	MYP 9	None	TMEM156 (1.00) KLHL5 (1.00)	0.500	0.151
rs2035383	MYP 9	APBB2	None	0.492	0.776
rs790142	MYP 9	APBB2	None	0.308	0.499
rs1565114	MYP 9	ATP8A1	None	0.331	0.920
rs1504491	MYP 9	None	GABRG1 (1.00)	0.500	0.370
rs225160	MYP 9	None	SPATA18 (1.00)	0.417	0.140
rs751266	MYP 9	FIP1L1	SCFD2 (1.00)	0.408	0.280

Table 7.1 Test of Imprinting Results (Step One): Transmission Distortion to Affected Offspring (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	p-value (z-statistic)
rs2538	MYP 9	None	CLOCK (1.00)	0.300	0.290
rs899631	MYP 9	POLR2B	IGFBP7 (1.00)	0.400	0.740
rs1456860	MYP 9	LPHN3	None	0.308	0.830
rs1879323	MYP 9	None	SRD5A2L2 (1.00)	0.433	0.130
rs1483720	MYP 9	SRD5A2L2	None	0.433	0.370
rs1899130	MYP 9	None	CENPC1 (1.00)	0.333	0.570
rs1560605	**MYP 9**	**None**	**SULT1B1 (1.00)**	**0.142**	**0.078 (-1.944)**
rs2063749	MYP 9	None	C4ORF7 (1.00) CSN3 (1.00)	0.325	0.999
rs9131	MYP 9	MTHFD2L	None	0.350	1.000
rs717239	MYP 9	FLJ25770	None	0.408	1.000
rs1511817	MYP 9	SHROOM3	None	0.276	0.252
rs1566485	MYP 9	SNOT6L	None	0.425	0.150
rs13429	MYP 10	None	C8ORF42 (1.00)	0.417	1.000
rs935559	MYP 10	C8ORF42	None	0.127	0.780
rs922798	MYP 10	CSMD1	None	0.433	0.920

Table 7.2 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Risk Group

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2 –value)	MAF	p-value (z-statistic)
rs770238	MYP 2	LPIN2	None	0.325	0.229
rs168206	MYP 2	DLGAP1	None	0.424	0.139
rs1565728	MYP 3	E2F7	None	0.475	0.054
rs998070	MYP 3	NAV3	None	0.483	0.219
rs2404772	MYP 3	None	CART1 (0.95) LRRIQ1 (1.00)	0.250	1.000
rs1508595	MYP 3	None	KITLG (1.00)	0.167	0.780
rs1401982	MYP 3	ATP2B1	None	0.417	0.007
rs1544921	MYP 3	CHPT1	SYCP3 (1.00) GNPTAB (1.00) FLJ11259 (0.88)	0.475	1.000
rs1922438	MYP 3	RFX4	None	0.475	0.450
rs2873108	MYP 4	DPP6	None	0.161	0.370
rs306278	MYP 4	DPP6	None	0.450	1.000
rs2033108	MYP 5	None	PCTP (1.00)	0.242	0.530
rs1024819	MYP 5	MSI2	None	0.492	0.510
rs1974692	MYP 5	MSI2	None	0.195	0.630
rs1881441	MYP 5	CLTC	None	0.117	0.450
rs1557720	MYP 5	BRIP1	None	0.458	0.410
rs715494	MYP 6	AP1B1	EWSR1 (0.95) GAS2L1 (0.95)	0.367	0.260
rs714027	MYP 6	HORMAD2	None	0.433	0.819
rs4444	MYP 6	OSBP2	None	0.442	0.092
rs762883	MYP 6	SYN3	None	0.400	1.000
rs9862	MYP 6	SYN3	None	0.467	1.000
rs739096	MYP 6	MYH9	None	0.458	1.000
rs2413411	MYP 6	CACNG2	None	0.325	0.826
rs760519	MYP 6	NCF4	FLJ90680 (0.94)	0.258	1.000
rs1534880	MYP 6	CSF2RB	None	0.492	1.000

Table 7.2 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Risk Group (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	p-value (z-statistic)
rs4348874	MYP 7	PTPNS	None	0.274	1.000
rs730348	MYP 7	NAV2	None	0.408	0.229
rs1470251	MYP 7	NAV2	None	0.183	0.600
rs1374719	MYP 7	SLC17A6	None	0.306	1.000
rs2928345	MYP 7	GAS2	None	0.192	0.490
rs1491846	MYP 7	None	KIF18A (0.81)	0.167	1.000
rs1032090	MYP 7	METT5D1	None	0.375	1.000
rs1564745	MYP 7	METT5D1	None	0.375	0.860
rs524373	MYP 7	None	KCNA4 (1.00)	0.317	0.410
rs2273544	MYP 7	TPC11L1	None	0.280	0.820
rs373499	MYP 7	CSTF3	None	0.423	1.000
rs1027695	MYP 8	None	ZIC4 (0.93)	0.494	0.811
rs765695	MYP 8	None	C3ORF58 (1.00)	0.442	0.170
rs723490	MYP 8	None	C3ORF58 (1.00)	0.460	0.167
rs1920395	MYP 8	P2RY14	None	0.325	0.802
rs755763	MYP 8	MBNL1	None	0.25	0.283
rs701265	MYP 8	P2RY1	None	0.217	1.000
rs9438	MYP 8	DHX36DEAH	None	0.350	1.000
rs1025192	MYP 8	MME	None	0.492	1.000
rs359573	MYP 8	None	PLCH1 (1.00)	0.307	0.098
rs986963	MYP 8	KCNAB1	None	0.317	1.000
rs1384542	MYP 8	FLJ16641	None	0.333	0.046
rs920417	MYP 8	None	SLITRK3 (1.00)	0.458	0.390
rs953834	MYP 8	None	GOLPH4 (1.00)	0.065	1.000
rs3863100	MYP 8	MDS1	None	0.075	1.000
rs905129	MYP 8	TNIK	None	0.442	0.170
rs1285082	MYP 8	FNDC313	None	0.458	0.300
rs2046718	MYP 8	NLGN1	None	0.308	1.000
rs753293	MYP 8	NAALADL2	None	0.267	1.000
rs1468924	MYP 8	KCNMB3	None	0.342	0.650
rs2049769	MYP 8	PEX5L	None	0.200	1.000
rs1973738	MYP 8	KLHL6	None	0.375	0.350
rs1401999	MYP 8	ABCC5	None	0.408	0.852
rs869417	MYP 8	ABCC5	None	0.408	0.852
rs4432622	MYP 8	VPS8	None	0.356	0.192
rs3332	MYP 8	VPS8	None	0.390	0.201
rs1837882	MYP 8	LIPH	None	0.458	1.000
rs6808013	MYP 8	DGKG	None	0.208	1.000

Table 7.2 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Risk Group (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2–value)	MAF	p-value (z-statistic)
rs1039559	MYP 9	TMEM156	None	0.458	0.116
rs974734	MYP 9	None	TMEM156 (1.00) KLHL5 (1.00)	0.500	0.042
rs2035383	MYP 9	APBB2	None	0.492	0.128
rs790142	MYP 9	APBB2	None	0.308	0.214
rs1565114	MYP 9	ATP8A1	None	0.331	0.690
rs1504491	MYP 9	None	GABRG1 (1.00)	0.500	0.820
rs225160	MYP 9	None	SPATA18 (1.00)	0.417	0.720
rs751266	MYP 9	FIP1L1	SCFD2 (1.00)	0.408	0.190
rs2538	MYP 9	None	CLOCK (1.00)	0.300	1.000
rs899631	MYP 9	POLR2B	IGFBP7 (1.00)	0.400	0.630
rs1456860	MYP 9	LPHN3	None	0.308	0.770
rs1879323	MYP 9	None	SRD5A2L2 (1.00)	0.433	0.130
rs13429	MYP 10	None	C8ORF42 (1.00)	0.417	0.845
rs935559	MYP 10	C8ORF42	None	0.127	1.000
rs922798	MYP 10	CSMD1	None	0.433	1.000

Table 7.3 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Non-Risk Group

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2–value)	MAF	p-value (z-statistic)
rs770238	MYP 2	LPIN2	None	0.325	0.192
rs168206	MYP 2	DLGAP1	None	0.424	0.358
rs1565728	MYP 3	E2F7	None	0.475	0.477
rs998070	MYP 3	NAV3	None	0.483	0.807
rs2404772	MYP 3	None	CART1 (0.95) LRRIQ1 (1.00)	0.250	0.782
rs1508595	MYP 3	None	KITLG (1.00)	0.167	0.423
rs1401982	MYP 3	ATP2B1	None	0.417	0.243
rs1544921	MYP 3	CHPT1	SYCP3 (1.00) GNPTAB (1.00) FLJ11259 (0.88)	0.475	0.906
rs1922438	MYP 3	RFX4	None	0.475	0.906
rs2873108	MYP 4	DPP6	None	0.161	0.642
rs306278	MYP 4	DPP6	None	0.450	0.617
rs2033108	MYP 5	None	PCTP (1.00)	0.242	0.661
rs1024819	MYP 5	MSI2	None	0.492	0.640
rs1974692	MYP 5	MSI2	None	0.195	0.764
rs1881441	MYP 5	CLTC	None	0.117	1.000
rs1557720	MYP 5	BRIP1	None	0.458	0.188
rs715494	**MYP 6**	**AP1B1**	**EWSR1 (0.95)** **GAS2L1 (0.95)**	**0.367**	**0.001 (5.054)** 0.660
rs714027	MYP 6	HORMAD2	None	0.433	0.300
rs4444	MYP 6	OSBP2	None	0.442	1.000
rs762883	MYP 6	SYN3	None	0.400	0.910
rs9862	MYP 6	SYN3	None	0.467	0.200
rs739096	MYP 6	MYH9	None	0.458	0.790
rs2413411	MYP 6	CACNG2	None	0.325	0.880
rs760519	MYP 6	NCF4	FLJ90680 (0.94)	0.258	0.130
rs1534880	MYP 6	CSF2RB	None	0.492	
rs4348874	MYP 7	PTPNS	None	0.274	0.770
rs730348	MYP 7	NAV2	None	0.408	0.520
rs1470251	MYP 7	NAV2	None	0.183	0.550
rs1374719	MYP 7	SLC17A6	None	0.306	0.590
rs2928345	MYP 7	GAS2	None	0.192	1.000
rs1491846	MYP 7	None	KIF18A (0.81)	0.167	0.870
rs1032090	MYP 7	METT5D1	None	0.375	0.280
rs1564745	MYP 7	METT5D1	None	0.375	0.280
rs524373	MYP 7	None	KCNA4 (1.00)	0.317	0.220
rs2273544	MYP 7	TPC11L1	None	0.280	1.000
rs373499	MYP 7	CSTF3	None	0.423	0.220

Table 7.3 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Non-Risk Group (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2-value)	MAF	p-value (z-statistic)
rs765695	MYP 8	None	C3ORF58 (1.00)	0.442	0.420
rs723490	MYP 8	None	C3ORF58 (1.00)	0.460	0.424
rs1027695	**MYP 8**	**None**	**ZIC4 (0.93)**	**0.494**	**0.021 (2.681)**
rs1920395	MYP 8	P2RY14	None	0.325	0.291
rs755763	MYP 8	MBNL1	None	0.25	0.488
rs701265	MYP 8	P2RY1	None	0.217	0.866
rs9438	MYP 8	DHX36DEAH	None	0.350	0.727
rs1025192	MYP 8	MME	None	0.492	0.355
rs359573	MYP 8	None	PLCH1 (1.00)	0.307	0.893
rs986963	MYP 8	KCNAB1	None	0.317	0.369
rs1384542	MYP 8	FLJ16641	None	0.333	0.338
rs920417	MYP 8	None	SLITRK3 (1.00)	0.458	0.901
rs953834	MYP 8	None	GOLPH4 (1.00)	0.065	0.674
rs3863100	MYP 8	MDS1	None	0.075	0.488
rs905129	**MYP 8**	**TNIK**	**None**	**0.442**	**0.039 (2.218)**
rs1285082	MYP 8	FNDC313	None	0.458	0.844
rs2046718	MYP 8	NLGN1	None	0.308	1.000
rs753293	MYP 8	NAALADL2	None	0.267	0.413
rs1468924	MYP 8	KCNMB3	None	0.342	0.595
rs2049769	**MYP 8**	**PEX5L**	**None**	**0.200**	**0.008 (3.090)**
rs1973738	MYP 8	KLHL6	None	0.375	0.246
rs1401999	MYP 8	ABCC5	None	0.408	0.750
rs869417	MYP 8	ABCC5	None	0.408	0.920
rs4432622	MYP 8	VPS8	None	0.356	0.680
rs3332	MYP 8	VPS8	None	0.390	1.000
rs1837882	MYP 8	LIPH	None	0.458	0.910
rs6808013	MYP 8	DGKG	None	0.208	1.000
rs1039559	MYP 9	TMEM156	None	0.458	0.620
rs974734	MYP 9	None	TMEM156 (1.00) KLHL5 (1.00)	0.500	1.000
rs2035383	MYP 9	APBB2	None	0.492	0.520
rs790142	MYP 9	APBB2	None	0.308	0.150
rs1565114	MYP 9	ATP8A1	None	0.331	1.000
rs1504491	MYP 9	None	GABRG1 (1.00)	0.500	0.270
rs225160	MYP 9	None	SPATA18 (1.00)	0.417	0.820
rs751266	MYP 9	FIP1L1	SCFD2 (1.00)	0.408	0.800
rs2538	MYP 9	None	CLOCK (1.00)	0.300	0.800
rs899631	MYP 9	POLR2B	IGFBP7 (1.00)	0.400	1.000

Table 7.3 Test of Imprinting Results (Step Two): SNP-allele count Asymmetry in Parents in the Non-Risk Group (Continuation)

SNP	MYP Region	Gene within which the chosen SNP is	Gene with which the chosen SNP is in high LD (r^2–value)	MAF	p-value (z-statistic)
rs1456860	MYP 9	LPHN3	None	0.308	0.790
rs1879323	MYP 9	None	SRD5A2L2 (1.00)	0.433	0.810
rs1483720	MYP 9	SRD5A2L2	None	0.433	1.000
rs1899130	MYP 9	None	CENPC1 (1.00)	0.333	0.290
rs1560605	MYP 9	None	SULT1B1 (1.00)	0.142	1.000
rs2063749	MYP 9	None	C4ORF7 (1.00) CSN3 (1.00)	0.325	0.780
rs9131	MYP 9	MTHFD2L	None	0.350	1.000
rs717239	MYP 9	FLJ25770	None	0.408	0.920
rs1511817	MYP 9	SHROOM3	None	0.276	0.680
rs1566485	MYP 9	SNOT6L	None	0.425	0.690
rs13429	MYP 10	None	C8ORF42 (1.00)	0.417	0.518
rs935559	**MYP 10**	**C8ORF42**	**None**	**0.127**	**0.079 (2.017)**
rs922798	MYP 10	CSMD1	None	0.433	0.476

7.4 Discussion

The effect of imprinting was examined in 91 trios with highly myopic offspring. This resulted in identification of 6 "offspring risk SNPs" and 4 "parental risk SNPs".

The first step of the TRIMM analysis addressed the issue of transmission distortion to affected offspring. With a borderline significant result, this test nominated a set of 6 "offspring risk SNPs" (rs1922438, rs4444, rs730348, rs753293, rs1837882, rs1560605) situated on 5 different chromosomes (Table 7.1). Some alleles of these SNPs were found to be over-transmitted (rs1922438, rs730348 and rs1837882), while others were under-transmitted, i.e. protective (rs4444, rs753293 and rs1560605) (Table 7.1). This finding fits well with the assumption of myopia being a complex disease with several loci affecting its susceptibility (the common disease, common variant theory). It is indeed possible that the over-transmission as well as simultaneous under-transmission of certain alleles would lead to myopia, explaining why the single candidate gene analyses carried out to date have been disappointing in identifying strong, reproducible genetic effects.

The second stage of the analysis was a test for imprinting, which comprised of two separate tests of SNP-count asymmetry among parents of highly myopic offspring: one performed in the group of trios whose offspring carried at least one copy of the set of risk alleles identified in step one (the risk group) and another in the group of trios whose offspring did not carry the nominated susceptibility alleles (the non-risk group). In the presence of an imprinting (parent-of-origin) effect, this test was expected to be significant in the risk group [379]. The results of this study, however, are equivocal, because the program produced two p-values for the overall test, one of which suggested statistical significance and the other which did not.

One of the calculated statistics, the z-score proved to be significant only in the non-risk group ($p=0.005$), showing no evidence ($p=0.256$) of SNP-count distortion in the parents of those offspring who carried at least one copy of the risk alleles identified in the first step of the study. One of the explanations for such a finding is that the nominated set of SNPs is protective and not disease causing (thus, a parental effect showed up only in the group with the opposite alleles- namely, those in the non-risk group).

Another explanation would be that there is a maternal effect rather than an imprinting effect. Maternal effects arise when the genetic and environmental characteristics of the mother influence the phenotype of her offspring, beyond the direct inheritance of alleles [374]. The mother plays crucial role not only as genetic parent, but also as a fetal environment. A maternal allele may, for example, damage a fetus through effects on the intrauterine milieu, regardless of whether the allele is passed to the offspring or not [380]. In the case of the imprinting, however, an allele must be transmitted to the offspring in order for it to exert its effect in an offspring. Therefore, finding a significant effect in the non-risk group only, may suggest the presence of maternal influence because the offspring of that group did not inherit the actual risk allele. In addition, the set of the "parental risk SNPs" identified as being "asymmetric" in parents differed from the set of the "offspring risk SNPs", again pointing towards a maternal effect, which is consistent with the absence of any transmission distortion to the affected offspring (in the case of imprinting the "offspring" and "parental" sets of risk SNPs should be identical or partly the same).

An additional test for SNP-allele count asymmetry in parents was also performed for the whole dataset without stratifying the trios. Reassuringly, this time both statistics (z-score and T^2) showed a highly significant result ($p=0.003$ for z-score and $p=0.006$ for T^2) and appointed a set of 8 "parental risk SNPs" (rs93559, rs770238, rs1401982, rs1557720, rs1384542, rs905129 and rs2049769). When comparing the 4 "parental risk SNPs" identified in the stratified analyses in the non-risk group (Table 7.3) and the 8 "parental risk SNPs" identified above, there were 3 SNPs common to both sets.

An asymmetry in SNP-counts between parents cannot by itself distinguish whether the mother or the father is responsible for the observed imbalance. However, because the only possible effect the father can have is via the genes he passes to the fetus (i.e. fetal effects) and there was no significant fetal effect (no transmission distortion) for the set of "parental risk SNPs", the data suggest that the set of "parental risk SNPs" is over-represented in the mothers and, thus, is exerting an effect via a maternally-mediated genetic influence.

When T^2-statistics were used to test for imprinting (instead of z-statistics) there was evidence of a parental SNP-allele count imbalance in both the risk group and the non-risk

group, suggesting the presence of a true imprinting (parent-of-origin) effect. Nonetheless, since the z-score revealed no statistical evidence in favour of such effect, it is difficult to confidently state that imprinting was present rather simply a maternal effect.

More generally, it should be noted that the present study was performed on a modestly sized sample and, thus, may have had insufficient power to detect imprinting (if present) with high confidence. In addition, a possible bias of selection might have been introduced because some trios included in these analyses were partial nuclear families from which an affected offspring was randomly chosen to form a trio with his/her parents, who in turn were selected based on their genotyping success as well as availability for recruitment.

The first step yielded a borderline significant set of "offspring risk SNPs", which makes it difficult to decide between a true and a false positive result. It would be desirable to perform these analyses on a larger dataset, that hopefully could reveal a more significant, convincing set of risk SNPs, leading to a more clearer identification o the presence or absence of the parent-of-origin effects.

7.5 Conclusion

In summary, the performed test for imprinting revealed an ambiguous result, leading to uncertainty whether or not myopia is affected by parent-of-origin effects and/or by maternal effects. Further analyses on a larger sample size are needed to resolve this question.

CHAPTER VIII.

GENERAL DISCUSSION AND FUTURE WORK

8.1 General Discussion

This study focused on the exploration of the genetic background of high myopia classified as equal or more severe than -6.00 D as this type of myopia threatens with permanent degradation of vision or even blindness.

All the analyses of this study were carried out on families with high myopia and cases/controls collected within the Family Study of Myopia. DNA samples were obtained in the form of posted mouthwashes.

As the source of DNA in this study was buccal cells, a series of tests were performed to ensure that DNA scheduled for further analyses was of good quality, to avoid unnecessary failure or false-positive genotyping. Firstly, the accuracy of human DNA quantification was assessed with four methods: spectrophotometry, fluorometry, gel electrophoresis and qPCR. Due to its specific primers, qPCR is the only approach that quantifies human DNA and, thus, it was considered as a reference in the experiment. In agreement with the literature, it was established that the traditional and most widely used method of spectrophotometry overestimated the amount of the human DNA by approximately 33% (compared to qPCR), but that fluorometry had the potential to substitute for human-specific qPCR, provided that the DNA sample was not degraded and was primarily of human origin. Nonetheless, spectrophotometry proved to be the most reproducible measure with the smallest coefficient of variability and, unlike fluorometry, it provided a measure of DNA purity (A_{260}/A_{280} ratio).

Further analyses performed on mouthwash-derived genetic material concerned the quality of DNA. The effect of lag time between mouthwash rinsing and the actual DNA extraction was examined, with regards to the quality of the extracted DNA. Given that mouthwashes were collected by post and assuming that the maximum mailing delay would be 3 days, mouthwashes collected for this experiment were processed on the same day or within one, two or three days of the mouth rinse procedure. DNA quality was evaluated with qPCR (as it provides information on the amount of human DNA in a sample and on the amplification ability of DNA needed for genotyping) and with agarose gel electrophoresis (as it gives insight into the degradation state of the DNA). Although no obvious

relationship between the amount of lag time and the presence/absence of degradation was revealed, it was established that, statistically, there was no effect of lag time on the quality of DNA, suggesting that mailing is an acceptable form of collection of mouthwash buccal cells. This observation was in agreement with previously published results of no such relationship [298].

In addition to the effect of lag time (up to 3 days) between mouthwash rinse and DNA extraction, the quality of DNA derived from mouthwashes collected for the study was also assessed in 500 subjects (1000 mouthwashes, as each participant provided two samples). As in the previous experiment, the quality of DNA was analyzed by gel electrophoresis and qPCR. To ensure that the quality of DNA screened by these two techniques was good enough for successful genotyping, a selection of those samples that proved to contain sufficient human DNA to provide robust amplification with qPCR and that were scored as non-degraded by gel electrophoresis were genotyped on an Illumina 6k Human bead array. A major finding from the experiment was that ~10% of DNA samples obtained from mouthwashes contained degraded DNA. Furthermore, there was an ~3-fold increased risk of DNA degradation in a participant's second mouthwash sample, given DNA degradation in their first, suggesting that DNA degradation may be due to factors specific to an individual subject. This finding may help the planning of mouthwash collection for large genome-wide analyses to be more cost-effective as it can be assumed beforehand that 10% of the samples may not be intact.

The average number of SNPs that could be genotyped on an Illumina array for each subject was 99.7% of the total, and the reliability of SNP genotyping "blind" duplicate mouthwash DNA samples was similarly high (>99.9% concordance). Nonetheless, it has to be mentioned that degraded samples were not sent out for genotyping and, thus, no comparison can be made between the genotyping success of "poor" and "good" quality of mouthwash-derived DNA.

Apart from investigating the quantity and quality of DNA extracted from mouthwashes, I also examined myopia candidate genes: analyses were performed to test for association

between myopia and polymorphisms in the myocilin gene (MYOC), the collagen type I alpha-1 gene (COL1A1), the collagen type II alpha-1 gene (COL2A1) and several genes located in MYP regions.

The myocilin gene is best known for its role in glaucoma [313, 314]. Genetic variants of MYOC, however, have also been implicated in causing susceptibility to high myopia [214, 321]. This study analyzed MYOC jointly with the research group of Prof. T.L.Young in Duke University (USA) and examined 250 nuclear families along with 112 highly myopic cases and 114 emmetropic controls. There was no significant heterogeneity in allele frequencies of genotyped variants between the families of USA and UK cohorts, or between founders of the families and cases/controls. Therefore, the pooling of subjects was "safe" from population stratification.

In contrast to the significant association between MYOC gene polymorphisms and high myopia found in Asian populations [214, 321], the Duke-Cardiff study suggested that there is no such relationship in subjects of Caucasian origin. Apart from the ethnic difference, another appealing explanation for this discrepancy is the smaller sample size and, thus, power of the studies examining Asian populations. Counter-intuitively, low-powered studies are more likely to give rise to false-positive associations than highly-powered ones. Furthermore, the MYOC gene variants which confer an increased risk of open angle glaucoma are different from those that may increase susceptibility to myopia. In this respect, the association of MYOC polymorphisms with both conditions may be coincidental.

MYP regions are chromosomal intervals linked to myopia of different grades and, consequently, genes within these loci are considered to be candidate genes for myopia susceptibility. Following this assumption, SNPs within those genes were tested for association with high grade myopia. Previously, several attempts have been made to find such a relationship and some have reported significant findings [179, 195, 207], but replication of most of these positive findings has failed [177, 185, 189, 193, 194, 199, 200, 204, 206, 208]. Continuing this line of disappointing replications, my study revealed no connection between the genes in MYP regions that were examined and high myopia.

There are several factors which could have been responsible for the negative result: modest sample size (low power), poor genome coverage (only 140 SNPs were examined) and different ethnicity to that of the original, positive studies (most of significant associations were found in Asian populations).

Excessive elongation of the eye is thought to be responsible for myopia development. Thus, collagen genes that occur in the eye's connective tissues (sclera and vitreous) were analyzed. Specific variants in the Collagen type I alpha-1 (COL1A1) and collagen type II alpha-1 (COL2A1) genes have been found to confer an increased risk to high myopia [163, 205]. Furthermore, both genes are responsible for connective tissue syndromes (Marfan and Stickler syndromes) with high myopia as a consistent phenotype. Nonetheless, this study found no convincing association between high myopia and the variants in COL1A1 or COL2A1 that were previously postulated to be myopia related.

Analysis of COL1A1 was performed on subjects of the same ethnicity as the original, positive study. The power of my study, however, was ~10% greater than that of the one with significant findings by Mutti et al [163]. However, Mutti et al defined myopia as <-0.75 D, while my investigation concentrated on high myopia only, classified as < -6.00 D. Thus, it is likely that COL1A1 is associated with lower degrees of myopia, or that the study of Mutti et al suffers from low power, which may have led to a false positive finding.

My analysis of the COL2A1 gene was performed on subjects of different ethnicity to that of the original, positive finding in a Japanese population [205] as the subjects examined here were of Caucasian origin. According to HapMap, the two populations (Caucasian and Japanese) exhibit dissimilar LD patterns in COL1A1, and the minor allele frequencies of the tested SNPs are significantly different. Aside from the discrepancies in ethnicities, the study of Inamori et al. [205] had lower power that that of the current study, suggesting that its positive finding may also represent a type I error. This assumption is supported by another two failed COL1A1 replications in Asian populations [169, 347].

In addition to the association analyses, this project examined the possibility of genetic imprinting in high myopia. It has already been observed that the correlation of refractive

error between siblings, and between mothers and offspring, is high, suggesting that the expression of only one parental allele (known as imprinting) may be behind this similarity. Driven by this idea, parents with an affected offspring (trios) were tested for allele-count disequilibrium. The results of this analysis did not convincingly support the presence of imprinting, but did show signs of a maternal effect. As the distinction between maternal effects and imprinting effects can be subtle (if present at all) and the analyzed sample size was rather modest (91 trios), it is not possible to state conclusively whether this study revealed any imprinting or maternal effect on high myopia, but the possibility is intriguing.

8.2 Future Work

The major drawback of this study is the lack of power due to the small sample size. Thus, one of the most important tasks for the future work would be to collect more participants to be able to perform more powerful and, consequently, more conclusive analyses. When my research project began, sample sizes of ~100 cases and 100 controls were considered large enough to allow the identification of susceptibility genes for complex disorders such as myopia. However, with the advent of genome-wide association studies researches now have realized that most genetic effects caused by commonly-occurring risk alleles have a much lower impact than was initially assumed, e.g. most risk-conferring alleles increase the chances of affectation by only ~20%. Rather than sample sizes of a few hundred participants, subject cohorts of several thousands are required to detect such genetic effects.

Recruitment of new subjects may target unrelated individuals (cases/controls) as well as families with highly myopic members. As it is still not clear whether association or linkage-like analyses would benefit myopia research best, collection of both types of subjects would be advantageous.

Another approach to increase the power of tests to dissect the genetic basis of high myopia would be to perform meta-analyses. The International Myopia Consortium has already

analyses on the existing collection of pedigrees.

REFERENCES

1. Goldschmidt, E., *[On the etiology of myopia. An epidemiological study.]*. Acta Ophthalmol (Copenh), 1968: p. Suppl 98:1+.
2. Dirani, M., et al., *Refractive errors in twin studies.* Twin Res Hum Genet, 2006. **9**(4): p. 566-72.
3. McCarty, C.A. and H.R. Taylor, *Myopia and vision 2020.* Am J Ophthalmol, 2000. **129**(4): p. 525-7.
4. Benjamin, W.J., *Borish's Clinical Refraction.* Second ed. 2006: Butterworth-Heinemann Elsevier.
5. Mouroulis, P., *Visual Instrumentation: Optical Design and Engineering principles.* 1999, New York: Mcgraw-Hill Publishing.
6. Young, T.L., R. Metlapally, and A.E. Shay, *Complex trait genetics of refractive error.* Arch Ophthalmol, 2007. **125**(1): p. 38-48.
7. Grosvenor, T.G., D, *Clinical Management of Myopia.* 1999: Butterworth-Heinemann.
8. Lo, P.I., et al., *Relationship between myopia and optical components--a study among Chinese Hong Kong student population.* Yan Ke Xue Bao, 1996. **12**(3): p. 121-5.
9. McBrien, N.A. and D.W. Adams, *A longitudinal investigation of adult-onset and adult-progression of myopia in an occupational group. Refractive and biometric findings.* Invest Ophthalmol Vis Sci, 1997. **38**(2): p. 321-33.
10. Olsen, T., et al., *On the ocular refractive components: the Reykjavik Eye Study.* Acta Ophthalmol Scand, 2007. **85**(4): p. 361-6.
11. Mutti, D.O., et al., *Axial growth and changes in lenticular and corneal power during emmetropization in infants.* Invest Ophthalmol Vis Sci, 2005. **46**(9): p. 3074-80.
12. Mutti, D.O., et al., *Refractive astigmatism and the toricity of ocular components in human infants.* Optom Vis Sci, 2004. **81**(10): p. 753-61.
13. Morgan, I.G., *The biological basis of myopic refractive error.* Clin Exp Optom, 2003. **86**(5): p. 276-88.
14. Troilo, D. and J. Wallman, *The regulation of eye growth and refractive state: an experimental study of emmetropization.* Vision Res, 1991. **31**(7-8): p. 1237-50.
15. Gwiazda, J., et al., *Emmetropization and the progression of manifest refraction in children followed from infancy to puberty.* Clinical vision sciences, 1993. **8**(4): p. 337-344.
16. Grosvenor, T. and D.A. Goss, *Role of the cornea in emmetropia and myopia.* Optom Vis Sci, 1998. **75**(2): p. 132-45.
17. Pennie, F.C., et al., *A longitudinal study of the biometric and refractive changes in full-term infants during the first year of life.* Vision Res, 2001. **41**(21): p. 2799-810.
18. Mutti, D.O., et al., *Optical and structural development of the crystalline lens in childhood.* Invest Ophthalmol Vis Sci, 1998. **39**(1): p. 120-33.
19. Friedman, N.E., D.O. Mutti, and K. Zadnik, *Corneal changes in schoolchildren.* Optom Vis Sci, 1996. **73**(8): p. 552-7.
20. Zadnik, K., et al., *Normal eye growth in emmetropic schoolchildren.* Optom Vis Sci, 2004. **81**(11): p. 819-28.

21. Zadnik, K., et al., *Longitudinal evidence of crystalline lens thinning in children.* Invest Ophthalmol Vis Sci, 1995. **36**(8): p. 1581-7.

22. Sorsby, A., et al., *Refraction and its components during the growth of the eye from the age of three.* Memo Med Res Counc, 1961. **301(Special)**(Special): p. 1-67.

23. Rymer, J. and C.F. Wildsoet, *The role of the retinal pigment epithelium in eye growth regulation and myopia: a review.* Vis Neurosci, 2005. **22**(3): p. 251-61.

24. Jacobi, Z., Broghammer and Pusch, *A genetic perspective on myopia.* Cell Mol Life Sci, 2005. **62**(7-8): p. 800-8.

25. Jones, L.A., et al., *Comparison of ocular component growth curves among refractive error groups in children.* Invest Ophthalmol Vis Sci, 2005. **46**(7): p. 2317-27.

26. Schaeffel, F., A. Glasser, and H.C. Howland, *Accommodation, refractive error and eye growth in chickens.* Vision Res, 1988. **28**(5): p. 639-57.

27. Wallman, J., J. Turkel, and J. Trachtman, *Extreme myopia produced by modest change in early visual experience.* Science, 1978. **201**(4362): p. 1249-51.

28. Mark, H.H., *Emmetropization. Physical aspects of a statistical phenomenon.* Ann Ophthalmol, 1972. **4**(5): p. 393-4 passim.

29. Goldschmidt, E., *The mystery of myopia.* Acta Ophthalmol Scand, 2003. **81**(5): p. 431-6.

30. Lyhne, N., et al., *The importance of genes and environment for ocular refraction and its determiners: a population based study among 20-45 year old twins.* Br J Ophthalmol, 2001. **85**(12): p. 1470-6.

31. Dirani, M., et al., *Heritability of refractive error and ocular biometrics: the Genes in Myopia (GEM) twin study.* Invest Ophthalmol Vis Sci, 2006. **47**(11): p. 4756-61.

32. Lin, L.L. and C.J. Chen, *Twin study on myopia.* Acta Genet Med Gemellol (Roma), 1987. **36**(4): p. 535-40.

33. Liang, C.L., et al., *Impact of family history of high myopia on level and onset of myopia.* Invest Ophthalmol Vis Sci, 2004. **45**(10): p. 3446-52.

34. Hammond, C.J., et al., *Genes and environment in refractive error: the twin eye study.* Invest Ophthalmol Vis Sci, 2001. **42**(6): p. 1232-6.

35. Sperduto, R.D., et al., *Prevalence of myopia in the United States.* Arch Ophthalmol, 1983. **101**(3): p. 405-7.

36. Crawford, H.E. and G.C. Hamman, *Racial analysis of ocular defects in the schools of Hawaii.* Hawaii Med J, 1949. **9**(2): p. 90-3.

37. Lin, L.L., et al., *Study of myopia among aboriginal school children in Taiwan.* Acta Ophthalmol Suppl, 1988. **185**: p. 34-6.

38. Bar Dayan, Y., et al., *The changing prevalence of myopia in young adults: a 13-year series of population-based prevalence surveys.* Invest Ophthalmol Vis Sci, 2005. **46**(8): p. 2760-5.

39. Lee, K.E., et al., *Changes in refraction over 10 years in an adult population: the Beaver Dam Eye study.* Invest Ophthalmol Vis Sci, 2002. **43**(8): p. 2566-71.

40. Matsumura, H. and H. Hirai, *Prevalence of myopia and refractive changes in students from 3 to 17 years of age.* Surv Ophthalmol, 1999. **44 Suppl 1**(1): p. S109-115.
41. Midelfart, A., et al., *Myopia among medical students in Norway.* Acta Ophthalmol (Copenh), 1992. **70**(3): p. 317-22.
42. Tay, M.T., et al., *Myopia and educational attainment in 421,116 young Singaporean males.* Ann Acad Med Singapore, 1992. **21**(6): p. 785-91.
43. Jiang, B.C., S. Schatz, and K. Seger, *Myopic progression and dark focus variation in optometric students during the first academic year.* Clin Exp Optom, 2005. **88**(3): p. 153-9.
44. Zylbermann, R., D. Landau, and D. Berson, *The influence of study habits on myopia in Jewish teenagers.* J Pediatr Ophthalmol Strabismus, 1993. **30**(5): p. 319-22.
45. Saw, S.M., et al., *Myopia: attempts to arrest progression.* Br J Ophthalmol, 2002. **86**(11): p. 1306-11.
46. von Noorden, G.K. and R.A. Lewis, *Ocular axial length in unilateral congenital cataracts and blepharoptosis.* Invest Ophthalmol Vis Sci, 1987. **28**(4): p. 750-2.
47. Wong, L., et al., *Education, reading, and familial tendency as risk factors for myopia in Hong Kong fishermen.* J Epidemiol Community Health, 1993. **47**(1): p. 50-3.
48. Parssinen, O., E. Hemminki, and A. Klemetti, *Effect of spectacle use and accommodation on myopic progression: final results of a three-year randomised clinical trial among schoolchildren.* Br J Ophthalmol, 1989. **73**(7): p. 547-51.
49. Cordain, L., et al., *An evolutionary analysis of the aetiology and pathogenesis of juvenile-onset myopia.* Acta Ophthalmol Scand, 2002. **80**(2): p. 125-35.
50. Edwards, M.H., *Do variations in normal nutrition play a role in the development of myopia?* Optom Vis Sci, 1996. **73**(10): p. 638-43.
51. Mandel, Y., et al., *Season of birth, natural light, and myopia.* Ophthalmology, 2008. **115**(4): p. 686-92.
52. McMahon, G., et al., *Season of birth, daylight hours at birth, and high myopia.* Ophthalmology, 2009. **116**(3): p. 468-73.
53. Rudnicka, A.R., et al., *Effect of breastfeeding and sociodemographic factors on visual outcome in childhood and adolescence.* Am J Clin Nutr, 2008. **87**(5): p. 1392-9.
54. Graham, M.V. and O.P. Gray, *Refraction of premature babies' eyes.* Br Med J, 1963. **1**(5343): p. 1452-4.
55. Quinn, G.E., et al., *Development of myopia in infants with birth weights less than 1251 grams. The Cryotherapy for Retinopathy of Prematurity Cooperative Group.* Ophthalmology, 1992. **99**(3): p. 329-40.
56. Rose, K.A., et al., *Outdoor activity reduces the prevalence of myopia in children.* Ophthalmology, 2008. **115**(8): p. 1279-85.
57. Grosvenor, T.G., D, *Clinical Management of Myopia.* 1999, Butterworth-Heinemann.
58. Weaver, H., *Genetics.* 1997: Wm. C. Brown Pblishers.

59. Mellor, J., *Dynamic nucleosomes and gene transcription.* Trends Genet, 2006. **22**(6): p. 320-9.
60. Wade, P.A., D. Pruss, and A.P. Wolffe, *Histone acetylation: chromatin in action.* Trends Biochem Sci, 1997. **22**(4): p. 128-32.
61. Yazgan, O. and J.E. Krebs, *Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes.* Biochem Cell Biol, 2007. **85**(4): p. 484-96.
62. Scherrer, K. and J. Jost, *The gene and the genon concept: a functional and information-theoretic analysis.* Mol Syst Biol, 2007. **3**: p. 87.
63. Weatherall, D.J., *Molecular pathology of single gene disorders.* J Clin Pathol, 1987. **40**(9): p. 959-70.
64. King, S., *A Dictionary of Genetics.* 2002: Oxford University Press.
65. Mattick, J.S., *Non-coding RNAs: the architects of eukaryotic complexity.* EMBO Rep, 2001. **2**(11): p. 986-91.
66. Mattick, J.S. and I.V. Makunin, *Non-coding RNA.* Hum Mol Genet, 2006. **15 Spec No 1**: p. R17-29.
67. Craig, J.M., *Heterochromatin--many flavours, common themes.* Bioessays, 2005. **27**(1): p. 17-28.
68. Grewal, S.I. and S. Jia, *Heterochromatin revisited.* Nat Rev Genet, 2007. **8**(1): p. 35-46.
69. Trojer, P. and D. Reinberg, *Facultative heterochromatin: is there a distinctive molecular signature?* Mol Cell, 2007. **28**(1): p. 1-13.
70. Harrison, P.R., *Molecular mechanisms involved in the regulation of gene expression during cell differentiation and development.* Immunol Ser, 1990. **49**: p. 411-64.
71. Tress, M.L., et al., *The implications of alternative splicing in the ENCODE protein complement.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5495-500.
72. Peter H. Raven, G.B.J., *Biology.* 6th ed. 2002: Boston: McGraw-Hill.
73. Villard, J., *Transcription regulation and human diseases.* Swiss Med Wkly, 2004. **134**(39-40): p. 571-9.
74. Cheung, V.G., et al., *Natural variation in human gene expression assessed in lymphoblastoid cells.* Nat Genet, 2003. **33**(3): p. 422-5.
75. Kamakaka, R.T., *Silencers and locus control regions: opposite sides of the same coin.* Trends Biochem Sci, 1997. **22**(4): p. 124-8.
76. Kuzmiak, H.A. and L.E. Maquat, *Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges.* Trends Mol Med, 2006. **12**(7): p. 306-16.
77. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.
78. Bernstein, B.E., A. Meissner, and E.S. Lander, *The mammalian epigenome.* Cell, 2007. **128**(4): p. 669-81.
79. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
80. Fazzari, M.J. and J.M. Greally, *Epigenomics: beyond CpG islands.* Nat Rev Genet, 2004. **5**(6): p.

446-55.

81. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.

82. Jones, P.A. and S.B. Baylin, *The fundamental role of epigenetic events in cancer.* Nat Rev Genet, 2002. **3**(6): p. 415-28.

83. Bartolomei, M.S. and S.M. Tilghman, *Genomic imprinting in mammals.* Annu Rev Genet, 1997. **31**: p. 493-525.

84. Luedi, P.P., et al., *Computational and experimental identification of novel human imprinted genes.* Genome Res, 2007. **17**(12): p. 1723-30.

85. Amor, D.J. and J. Halliday, *A review of known imprinting syndromes and their association with assisted reproduction technologies.* Hum Reprod, 2008. **23**(12): p. 2826-34.

86. Reich, D.E., et al., *Human genome sequence variation and the influence of gene history, mutation and recombination.* Nat Genet, 2002. **32**(1): p. 135-42.

87. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations.* Science, 2005. **307**(5712): p. 1072-9.

88. Schneider, J.A., et al., *DNA variability of human genes.* Mech Ageing Dev, 2003. **124**(1): p. 17-25.

89. Salisbury, B.A., et al., *SNP and haplotype variation in the human genome.* Mutat Res, 2003. **526**(1-2): p. 53-61.

90. Elahi, E., J. Kumm, and M. Ronaghi, *Global genetic analysis.* J Biochem Mol Biol, 2004. **37**(1): p. 11-27.

91. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

92. Subramanian, S., R.K. Mishra, and L. Singh, *Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions.* Genome Biol, 2003. **4**(2): p. R13.

93. Katti, M.V., P.K. Ranjekar, and V.S. Gupta, *Differential distribution of simple sequence repeats in eukaryotic genome sequences.* Mol Biol Evol, 2001. **18**(7): p. 1161-7.

94. Li, Y.C., et al., *Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review.* Mol Ecol, 2002. **11**(12): p. 2453-65.

95. Li, Y.C., et al., *Microsatellites within genes: structure, function, and evolution.* Mol Biol Evol, 2004. **21**(6): p. 991-1007.

96. Hui, J., et al., *Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing.* Embo J, 2005. **24**(11): p. 1988-98.

97. Weber, J.L. and P.E. May, *Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction.* Am J Hum Genet, 1989. **44**(3): p. 388-96.

98. Chu, C.S., et al., *Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA.* Nat Genet, 1993. **3**(2): p. 151-6.

99. Cuppens, H., et al., *Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a*

disease mutation. J Clin Invest, 1998. **101**(2): p. 487-96.

100. Buerger, H., et al., *Allelic length of a CA dinucleotide repeat in the egfr gene correlates with the frequency of amplifications of this sequence--first results of an inter-ethnic breast cancer study.* J Pathol, 2004. **203**(1): p. 545-50.

101. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nat Genet, 1999. **22**(3): p. 231-8.

102. Consortium, T.I.H., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

103. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.

104. Beckmann, J.S., X. Estivill, and S.E. Antonarakis, *Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability.* Nat Rev Genet, 2007. **8**(8): p. 639-46.

105. Ionita-Laza, I., et al., *Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis.* Genomics, 2009. **93**(1): p. 22-6.

106. Mayeux, R., *Mapping the new frontier: complex genetic disorders.* J Clin Invest, 2005. **115**(6): p. 1404-7.

107. Wojczynski, M.K. and H.K. Tiwari, *Definition of phenotype.* Adv Genet, 2008. **60**: p. 75-105.

108. Schulze, T.G. and F.J. McMahon, *Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping.* Hum Hered, 2004. **58**(3-4): p. 131-8.

109. Wang, Y., R. Ottman, and D. Rabinowitz, *A method for estimating penetrance from families sampled for linkage analysis.* Biometrics, 2006. **62**(4): p. 1081-8.

110. Ravine, D. and D.N. Cooper, *Adult-onset genetic disease: mechanisms, analysis and prediction.* QJM, 1997. **90**(2): p. 83-103.

111. Schork, N.J., *Genetics of complex disease: approaches, problems, and solutions.* Am J Respir Crit Care Med, 1997. **156**(4 Pt 2): p. S103-9.

112. Evans, D.G. and R. Harris, *Heterogeneity in genetic conditions.* Q J Med, 1992. **84**(304): p. 563-5.

113. Glazier, A.M., J.H. Nadeau, and T.J. Aitman, *Finding genes that underlie complex traits.* Science, 2002. **298**(5602): p. 2345-9.

114. Nadeau, J.H., *Modifier genes in mice and humans.* Nat Rev Genet, 2001. **2**(3): p. 165-74.

115. Fearnhead, N.S., B. Winney, and W.F. Bodmer, *Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model.* Cell Cycle, 2005. **4**(4): p. 521-5.

116. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.

117. Iyengar, S.K. and R.C. Elston, *The genetic basis of complex traits: rare variants or "common gene, common disease"?* Methods Mol Biol, 2007. **376**: p. 71-84.

118. Schafer, A.J. and J.R. Hawkins, *DNA variation and the future of human genetics.* Nat Biotechnol, 1998. **16**(1): p. 33-9.

119. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.

120. Cain, M., *Discover Biology*. 3rd Edition ed. 2006: W W Norton & Co Ltd.

121. Pritchard, J.K. and M. Przeworski, *Linkage disequilibrium in humans: models and data.* Am J Hum Genet, 2001. **69**(1): p. 1-14.

122. Zondervan, K.T. and L.R. Cardon, *The complex interplay among factors that influence allelic association.* Nat Rev Genet, 2004. **5**(2): p. 89-100.

123. Consortium, I.H., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.

124. Sebastiani, P., et al., *Minimal haplotype tagging.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9900-5.

125. Devlin, B. and N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping.* Genomics, 1995. **29**(2): p. 311-22.

126. Lewontin, R.C., *The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models.* Genetics, 1964. **49**(1): p. 49-67.

127. Balding, D.J., *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. 7(10): p. 781-91.

128. Lewontin, R.C., *On measures of gametic disequilibrium.* Genetics, 1988. **120**(3): p. 849-52.

129. Hardy, G.H., *MENDELIAN PROPORTIONS IN A MIXED POPULATION.* Science, 1908. **28**(706): p. 49-50.

130. Trikalinos, T.A., et al., *Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations.* Am J Epidemiol, 2006. **163**(4): p. 300-9.

131. Leal, S.M., *Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium.* Genet Epidemiol, 2005. **29**(3): p. 204-14.

132. Wittke-Thompson, J.K., A. Pluzhnikov, and N.J. Cox, *Rational inferences about departures from Hardy-Weinberg equilibrium.* Am J Hum Genet, 2005. **76**(6): p. 967-86.

133. Elston, R.C., *Testing for Hardy-Weinberg Equilibrium in Small Samples.* Biometrics, 1977: p. 536-41.

134. Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium.* Am J Hum Genet, 2005. **76**(5): p. 887-93.

135. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* Am J Hum Genet, 1993. **52**(3): p. 506-16.

136. Patil, N., et al., *Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.* Science, 2001. **294**(5547): p. 1719-23.

137. Gabriel, S.B., et al., *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.

138. Johnson, G.C., et al., *Haplotype tagging for the identification of common disease genes.* Nat Genet, 2001. **29**(2): p. 233-7.

139. Laird, N.M. and C. Lange, *Family-based designs in the age of large-scale gene-association studies.* Nat Rev Genet, 2006. 7(5): p. 385-94.

140. Schulze, T.G. and F.J. McMahon, *Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines.* Am J Med Genet, 2002. **114**(1): p. 1-11.

141. Falk, C.T. and P. Rubinstein, *Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations.* Ann Hum Genet, 1987. **51**(Pt 3): p. 227-33.

142. Spielman, R.S. and W.J. Ewens, *The TDT and other family-based tests for linkage disequilibrium and association.* Am J Hum Genet, 1996. **59**(5): p. 983-9.

143. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.

144. Pritchard, J.K., et al., *Association mapping in structured populations.* Am J Hum Genet, 2000. **67**(1): p. 170-81.

145. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

146. Hirschhorn, J.N., et al., *A comprehensive review of genetic association studies.* Genet Med, 2002. **4**(2): p. 45-61.

147. Dudbridge, F., *Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data.* Hum Hered, 2008. **66**(2): p. 87-98.

148. Spielman, R.S. and W.J. Ewens, *A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test.* Am J Hum Genet, 1998. **62**(2): p. 450-8.

149. Martin, E.R., et al., *Accounting for linkage in family-based tests of association with missing parental genotypes.* Am J Hum Genet, 2003. **73**(5): p. 1016-26.

150. McGinnis, R., S. Shifman, and A. Darvasi, *Power and efficiency of the TDT and case-control design for association scans.* Behav Genet, 2002. **32**(2): p. 135-44.

151. Wang, W.Y., et al., *Genome-wide association studies: theoretical and practical concerns.* Nat Rev Genet, 2005. **6**(2): p. 109-18.

152. Chapman, J.M., et al., *Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power.* Hum Hered, 2003. **56**(1-3): p. 18-31.

153. Pennacchio, L.A. and E.M. Rubin, *Genomic strategies to identify mammalian regulatory sequences.* Nat Rev Genet, 2001. **2**(2): p. 100-9.

154. Huizinga, T.W., D.S. Pisetsky, and R.P. Kimberly, *Associations, populations, and the truth: recommendations for genetic association studies in Arthritis & Rheumatism.* Arthritis Rheum, 2004. **50**(7): p. 2066-71.

155. Newton-Cheh, C. and J.N. Hirschhorn, *Genetic association studies of complex traits: design and analysis issues.* Mutat Res, 2005. **573**(1-2): p. 54-69.

156. Goldschmidt, E., *The importance of heredity and environment in the etiology of low myopia.* Acta Ophthalmol (Copenh), 1981. **59**(5): p. 759-62.

157. Goss, D.A. and T.W. Jackson, *Clinical findings before the onset of myopia in youth: 4. Parental history of myopia.* Optom Vis Sci, 1996. **73**(4): p. 279-82.

158. Zadnik, K., et al., *The effect of parental history of myopia on children's eye size.* Jama, 1994.

271(17): p. 1323-7.

159. Yap, M., et al., *Role of heredity in the genesis of myopia.* Ophthalmic Physiol Opt, 1993. **13**(3): p. 316-9.

160. Young, T.L., et al., *A second locus for familial high myopia maps to chromosome 12q.* Am J Hum Genet, 1998. **63**(5): p. 1419-24.

161. Li, Y.J., et al., *An International Collaborative Family-based Whole Genome Linkage Scan for High-grade Myopia.* Invest Ophthalmol Vis Sci, 2009.

162. Guggenheim, J.A., G. Kirov, and S.A. Hodson, *The heritability of high myopia: a reanalysis of Goldschmidt's data.* J Med Genet, 2000. **37**(3): p. 227-31.

163. Mutti, D.O., et al., *Candidate gene and locus analysis of myopia.* Mol Vis, 2007. **13**: p. 1012-9.

164. Li, J., et al., *[The SNPs analysis of encoding sequence of interacting factor gene in Chinese population].* Zhonghua Yi Xue Yi Chuan Xue Za Zhi, 2003. **20**(5): p. 454-6.

165. Scavello, G.S., et al., *Sequence variants in the transforming growth beta-induced factor (TGIF) gene are not associated with high myopia.* Invest Ophthalmol Vis Sci, 2004. **45**(7): p. 2091-7.

166. Paluru, P.C., et al., *Exclusion of lumican and fibromodulin as candidate genes in MYP3 linked high grade myopia.* Mol Vis, 2004. **10**: p. 917-22.

167. Hasumi, Y., et al., *Analysis of single nucleotide polymorphisms at 13 loci within the transforming growth factor-induced factor gene shows no association with high myopia in Japanese subjects.* Immunogenetics, 2006. **58**(12): p. 947-53.

168. Simpson, C.L., et al., *The Roles of PAX6 and SOX2 in Myopia: Lessons from the 1958 British Birth Cohort.* Invest Ophthalmol Vis Sci, 2007. **48**(10): p. 4421-5.

169. Liang, C.L., et al., *Systematic assessment of the tagging polymorphisms of the COL1A1 gene for high myopia.* J Hum Genet, 2007. **52**(4): p. 374-7.

170. Schwartz, M., M. Haim, and D. Skarsholm, *X-linked myopia: Bornholm eye disease. Linkage to DNA markers on the distal part of Xq.* Clin Genet, 1990. **38**(4): p. 281-6.

171. Haim, M., H.C. Fledelius, and Skarsholm, *X-linked myopia in Danish family.* Acta Ophthalmol (Copenh), 1988. **66**(4): p. 450-6.

172. Young, T.L., et al., *Evidence that a locus for familial high myopia maps to chromosome 18p.* Am J Hum Genet, 1998. **63**(1): p. 109-19.

173. Farbrother, J.E., et al., *Linkage analysis of the genetic loci for high myopia on 18p, 12q, and 17q in 51 U.K. families.* Invest Ophthalmol Vis Sci, 2004. **45**(9): p. 2879-85.

174. Ibay, G., et al., *Candidate high myopia loci on chromosomes 18p and 12q do not play a major role in susceptibility to common myopia.* BMC Med Genet, 2004. **5**(20): p. 20.

175. Lam, D.S., et al., *Familial high myopia linkage to chromosome 18p.* Ophthalmologica, 2003. **217**(2): p. 115-8.

176. Heath, S., et al., *A novel approach to search for identity by descent in small samples of patients and controls from the same mendelian breeding unit: a pilot study on myopia.* Hum Hered, 2001. **52**(4): p. 183-90.

177. Young, T.L., *Molecular genetics of human myopia: an update.* Optom Vis Sci, 2009. **86**(1): p. E8-E22.

178. Stambolian, D., et al., *Genomewide linkage scan for myopia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 22q12.* Am J Hum Genet, 2004. **75**(3): p. 448-59.

179. Nurnberg, G., et al., *Refinement of the MYP3 locus on human chromosome 12 in a German family with Mendelian autosomal dominant high-grade myopia by SNP array mapping.* Int J Mol Med, 2008. **21**(4): p. 429-38.

180. Naiglin, L., et al., *A genome wide scan for familial high myopia suggests a novel locus on chromosome 7q36.* J Med Genet, 2002. **39**(2): p. 118-24.

181. Paluru, P., et al., *New locus for autosomal dominant high myopia maps to the long arm of chromosome 17.* Invest Ophthalmol Vis Sci, 2003. **44**(5): p. 1830-6.

182. Stambolian, D., et al., *Genome-wide scan of additional Jewish families confirms linkage of a myopia susceptibility locus to chromosome 22q12.* Mol Vis, 2006. **12**: p. 1499-505.

183. Klein, A.P., et al., *Confirmation of linkage to ocular refraction on chromosome 22q and identification of a novel linkage region on 1q.* Arch Ophthalmol, 2007. **125**(1): p. 80-5.

184. Stambolian, D., et al., *Genome-wide scan for myopia in the Old Order Amish.* Am J Ophthalmol, 2005. **140**(3): p. 469-76.

185. Hammond, C.J., et al., *A susceptibility locus for myopia in the normal population is linked to the PAX6 gene region on chromosome 11: a genomewide scan of dizygotic twins.* Am J Hum Genet, 2004. **75**(2): p. 294-304.

186. Andrew, T., et al., *Identification and replication of three novel myopia common susceptibility gene loci on chromosome 3q26 using linkage and linkage disequilibrium mapping.* PLoS genetics, 2008. **4**(10): p. e1000220.

187. Zhang, Q., et al., *A new locus for autosomal dominant high myopia maps to 4q22-q27 between D4S1578 and D4S1612.* Mol Vis, 2005. **11**: p. 554-60.

188. Paluru, P.C., et al., *Identification of a novel locus on 2q for autosomal dominant high-grade myopia.* Invest Ophthalmol Vis Sci, 2005. **46**(7): p. 2300-7.

189. Chen, C.Y., et al., *Linkage replication of the MYP12 locus in common myopia.* Invest Ophthalmol Vis Sci, 2007. **48**(10): p. 4433-9.

190. Zhang, Q., et al., *Novel locus for X linked recessive high myopia maps to Xq23-q25 but outside MYP1.* J Med Genet, 2006. **43**(5): p. e20.

191. Zhang, Q., et al., *Confirmation of a genetic locus for X-linked recessive high myopia outside MYP1.* J Hum Genet, 2007. **52**(5): p. 469-72.

192. Wojciechowski, R., et al., *Genomewide scan in Ashkenazi Jewish families demonstrates evidence of linkage of ocular refraction to a QTL on chromosome 1p36.* Hum Genet, 2006. **119**(4): p. 389-99.

193. Lam, C.Y., et al., *A genome-wide scan maps a novel high myopia locus to 5p15.* Invest Ophthalmol Vis Sci, 2008. **49**(9): p. 3768-78.

194. Ciner, E., et al., *Genomewide scan of ocular refraction in African-American families shows significant linkage to chromosome 7p15.* Genet Epidemiol, 2008. **32**(5): p. 454-63.

195. Metlapally, R., et al., *Evaluation of the X-Linked High Grade Myopia Locus (MYP1) with Cone Dysfunction and Color Vision Deficiencies.* Invest Ophthalmol Vis Sci, 2008.

196. Zhang, Q., et al., *Mutations in NYX of individuals with high myopia, but without night blindness.* Mol Vis, 2007. **13**: p. 330-6.

197. Lam, D.S., et al., *TGFbeta-induced factor: a candidate gene for high myopia.* Invest Ophthalmol Vis Sci, 2003. **44**(3): p. 1012-5.

198. Wang, P., et al., *High myopia is not associated with the SNPs in the TGIF, Lumican, TGFB1, and HGF genes.* Invest Ophthalmol Vis Sci, 2008.

199. Pertile, K.K., et al., *Assessment of TGIF as a candidate gene for myopia.* Invest Ophthalmol Vis Sci, 2008. **49**(1): p. 49-54.

200. Scavello, G.S., Jr., et al., *Genomic structure and organization of the high grade Myopia-2 locus (MYP2) critical region: mutation screening of 9 positional candidate genes.* Mol Vis, 2005. **11**: p. 97-110.

201. Young, T.L., *Dissecting the genetics of human high myopia: a molecular biologic approach.* Trans Am Ophthalmol Soc, 2004. **102**: p. 423-45.

202. Zhou, J. and T.L. Young, *Evaluation of Lipin 2 as a candidate gene for autosomal dominant 1 high-grade myopia.* Gene, 2005. **352**: p. 10-9.

203. Wang, I.J., et al., *The association of single nucleotide polymorphisms in the 5'-regulatory region of the lumican gene with susceptibility to high myopia in Taiwan.* Mol Vis, 2006. **12**: p. 852-7.

204. Majava, M., et al., *Novel mutations in the small leucine-rich repeat protein/proteoglycan (SLRP) genes in high myopia.* Hum Mutat, 2007. **28**(4): p. 336-44.

205. Inamori, Y., et al., *The COL1A1 gene and high myopia susceptibility in Japanese.* Hum Genet, 2007. **122**(2): p. 151-7.

206. Han, W., et al., *Association of PAX6 polymorphisms with high myopia in Han Chinese nuclear families.* Invest Ophthalmol Vis Sci, 2009. **50**(1): p. 47-56.

207. Tsai, Y.Y., et al., *A PAX6 gene polymorphism is associated with genetic predisposition to extreme myopia.* Eye, 2008. **22**(4): p. 576-81.

208. Hewitt, A.W., et al., *PAX6 mutations may be associated with high myopia.* Ophthalmic Genet, 2007. **28**(3): p. 179-82.

209. Hayashi, T., et al., *Exclusion of transforming growth factor-beta1 as a candidate gene for myopia in the Japanese.* Jpn J Ophthalmol, 2007. **51**(2): p. 96-9.

210. Lin, H.J., et al., *The TGFbeta1 gene codon 10 polymorphism contributes to the genetic predisposition to high myopia.* Mol Vis, 2006. **12**: p. 698-703.

211. Han, W., et al., *Family-based association analysis of hepatocyte growth factor (HGF) gene polymorphisms in high myopia.* Invest Ophthalmol Vis Sci, 2006. **47**(6): p. 2291-9.

212. Liang, C.L., et al., *Evaluation of MMP3 and TIMP1 as candidate genes for high myopia in young*

Taiwanese men. Am J Ophthalmol, 2006. **142**(3): p. 518-20.

213. Leung, Y.F., et al., *TIGR/MYOC proximal promoter GT-repeat polymorphism is not associated with myopia.* Hum Mutat, 2000. **16**(6): p. 533.

214. Tang, W.C., et al., *Linkage and association of myocilin (MYOC) polymorphisms with high myopia in a Chinese population.* Molecular Vision, 2007. **13**(57-61): p. 534-544.

215. Wu, H., Yu, X and Yap E, *Allelic association between trabecular meshwork-induced glucocorticoid response (TIGR) gene and sporadic severe myopia.* Invest Ophthalmol Vis Sci, 1999: p. p.S600, Abstract 3148.

216. Li, T., et al., *Evaluation of EGR1 as a candidate gene for high myopia.* Mol Vis, 2008. **14**: p. 1309-12.

217. Feigelson, H.S., et al., *Determinants of DNA yield and quality from buccal cell samples collected with mouthwash.* Cancer Epidemiol Biomarkers Prev, 2001. **10**(9): p. 1005-8.

218. Singer, V.L., et al., *Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation.* Anal Biochem, 1997. **249**(2): p. 228-38.

219. Blotta, I., et al., *Quantitative assay of total dsDNA with PicoGreen reagent and real-time fluorescent detection.* Ann Ist Super Sanita, 2005. **41**(1): p. 119-23.

220. Vitzthum, F., et al., *A quantitative fluorescence-based microplate assay for the determination of double-stranded DNA using SYBR Green I and a standard ultraviolet transilluminator gel imaging system.* Anal Biochem, 1999. **276**(1): p. 59-64.

221. Rengarajan, K., et al., *Quantifying DNA concentrations using fluorometry: a comparison of fluorophores.* Mol Vis, 2002. **8**: p. 416-21.

222. J. Sambrook, E.F.F., T. Maniatis, *Molecular Cloning: A laboratory manual.* 1989: Cold Spring Harbor Laboratory Press.

223. Sambrook, J. and D. Russell, *Molecular Cloning: a laboratory manual.* 3rd ed. 2001, New York: Cold Spring Harbor Laboratory Press.

224. Wilfinger, W.W., K. Mackey, and P. Chomczynski, *Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity.* Biotechniques, 1997. **22**(3): p. 474-6, 478-81.

225. Kim, H.S., S.H. Byun, and B.M. Lee, *Effects of chemical carcinogens and physicochemical factors on the UV spectrophotometric determination of DNA.* J Toxicol Environ Health A, 2005. **68**(23-24): p. 2081-95.

226. Valeur, B., *Molecular Fluorescence: Principles and Applications.* 2001: Wiley-VCH Verlag GmbH.

227. Sambrook, F., Maniatis, *Molecular Cloning: A laboratory manual.* 1989: Cold Spring Harbor Laboratory Press.

228. Le Pecq, J.B. and C. Paoletti, *A new fluorometric method for RNA and DNA determination.* Anal Biochem, 1966. **17**(1): p. 100-7.

229. Ahn, S.J., J. Costa, and J.R. Emanuel, *PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR.* Nucleic Acids Res, 1996. **24**(13): p. 2623-5.

230. Schneeberger, C., et al., *Quantitative detection of reverse transcriptase-PCR products by means of a novel and sensitive DNA stain.* PCR Methods Appl, 1995. **4**(4): p. 234-8.

231. Tuma, R.S., et al., *Characterization of SYBR Gold nucleic acid gel stain: a dye optimized for use with 300-nm ultraviolet transilluminators.* Anal Biochem, 1999. **268**(2): p. 278-88.

232. Martin, R., *Gel Electrophoresis: Nucleic Acids.* 1996: BIOS Scientific Publishers Limited.

233. Voytas, D., *Agarose gel electrophoresis.* Curr Protoc Immunol, 2001. **Chapter 10**: p. Unit 10 4.

234. Huang, Q. and W.L. Fu, *Comparative analysis of the DNA staining efficiencies of different fluorescent dyes in preparative agarose gel electrophoresis.* Clin Chem Lab Med, 2005. **43**(8): p. 841-2.

235. Gottlieb, M. and M. Chavko, *Silver staining of native and denatured eucaryotic DNA in agarose gels.* Anal Biochem, 1987. **165**(1): p. 33-7.

236. Peats, S., *Quantitation of protein and DNA in silver-stained agarose gels.* Anal Biochem, 1984. **140**(1): p. 178-82.

237. Raymer, D.M. and D.E. Smith, *A simple system for staining protein and nucleic acid electrophoresis gels.* Electrophoresis, 2007. **28**(5): p. 746-8.

238. Stothard, J.R., I.A. Frame, and M.A. Miles, *An evaluation of four staining methods for the detection of DNA in nondenaturing polyacrylamide gels.* Anal Biochem, 1997. **253**(2): p. 262-4.

239. Jin, L.T. and J.K. Choi, *Usefulness of visible dyes for the staining of protein or DNA in electrophoresis.* Electrophoresis, 2004. **25**(15): p. 2429-38.

240. White, H.W. and M. Wu, *Factors affecting quantitation of DNA bands in gels using a charge-coupled device imaging system.* Electrophoresis, 2001. **22**(5): p. 860-3.

241. Mullis, K., et al., *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction.* Cold Spring Harb Symp Quant Biol, 1986. **51 Pt 1**: p. 263-73.

242. Strachan, R., *Human molecular genetics 2.* 1999: BIOS Scuentific Publishers Ltd.

243. Kubista, M., et al., *The real-time polymerase chain reaction.* Mol Aspects Med, 2006. **27**(2-3): p. 95-125.

244. Higuchi, R., et al., *Simultaneous amplification and detection of specific DNA sequences.* Biotechnology (N Y), 1992. **10**(4): p. 413-7.

245. Ririe, K.M., R.P. Rasmussen, and C.T. Wittwer, *Product differentiation by analysis of DNA melting curves during the polymerase chain reaction.* Anal Biochem, 1997. **245**(2): p. 154-60.

246. Wilhelm, J. and A. Pingoud, *Real-time polymerase chain reaction.* Chembiochem, 2003. **4**(11): p. 1120-8.

247. Rutledge, R.G. and C. Cote, *Mathematics of quantitative kinetic PCR and the application of standard curves.* Nucleic Acids Res, 2003. **31**(16): p. e93.

248. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.* Science, 1988. **239**(4839): p. 487-91.

249. Higuchi, R., et al., *Kinetic PCR analysis: real-time monitoring of DNA amplification reactions.* Biotechnology (N Y), 1993. **11**(9): p. 1026-30.

250. Bio-Rad, L., *Real-Time PCR Applications Guide*. 2006.

251. Bar, T., et al., *Kinetic Outlier Detection (KOD) in real-time PCR.* Nucleic Acids Res, 2003. **31**(17): p. e105.

252. Rossen, L., et al., *Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions.* Int J Food Microbiol, 1992. **17**(1): p. 37-45.

253. Jorgenson, J.W. and K.D. Lukacs, *Free-zone electrophoresis in glass capillaries.* Clin Chem, 1981. **27**(9): p. 1551-3.

254. Wenz, H., et al., *High-precision genotyping by denaturing capillary electrophoresis.* Genome Res, 1998. **8**(1): p. 69-80.

255. Holland, P.M., et al., *Detection of specific polymerase chain reaction product by utilizing the 5'---3' exonuclease activity of Thermus aquaticus DNA polymerase.* Proc Natl Acad Sci U S A, 1991. **88**(16): p. 7276-80.

256. Tyagi, S. and F.R. Kramer, *Molecular beacons: probes that fluoresce upon hybridization.* Nat Biotechnol, 1996. **14**(3): p. 303-8.

257. Svanvik, N., et al., *Detection of PCR products in real time using light-up probes.* Anal Biochem, 2000. **287**(1): p. 179-82.

258. Schmerer, W.M., S. Hummel, and B. Herrmann, *[Reproducibility of aDNA typing].* Anthropol Anz, 1997. **55**(2): p. 199-206.

259. Taberlet, P., et al., *Reliable genotyping of samples with very low DNA quantities using PCR.* Nucleic Acids Res, 1996. **24**(16): p. 3189-94.

260. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

261. Lindsey, J.K., *Construction and Comparison of Statistical Models.* Journal of the Royal Statistical Society. Series B (methological). 1973. **36**(3): p. 418-425.

262. Kao, L.S. and C.E. Green, *Analysis of variance: is there a difference in means and what does it mean?* J Surg Res, 2008. **144**(1): p. 158-70.

263. Bewick, V., L. Cheek, and J. Ball, *Statistics review 9: one-way analysis of variance.* Crit Care, 2004. **8**(2): p. 130-6.

264. Zou, K.H., K. Tuncali, and S.G. Silverman, *Correlation and simple linear regression.* Radiology, 2003. **227**(3): p. 617-22.

265. Twomey, P.J. and M.H. Kroll, *How to use linear regression and correlation in quantitative method comparison studies.* Int J Clin Pract, 2008. **62**(4): p. 529-38.

266. Bewick, V., L. Cheek, and J. Ball, *Statistics review 7: Correlation and regression.* Crit Care, 2003. **7**(6): p. 451-9.

267. Bewick, V., L. Cheek, and J. Ball, *Statistics review 8: Qualitative data - tests of association.* Crit Care, 2004. **8**(1): p. 46-53.

268. Sistrom, C.L. and C.W. Garvan, *Proportions, odds, and risk.* Radiology, 2004. **230**(1): p. 12-9.
269. Kirkwood, B.R. and J.A.C. Sterne, *Essential Medical Statistics.* Second Edition ed. 2003: Blackwell Publishing.
270. Altman, D.G., *Practical statistics for medical research.* 1991: Chapman&Hall.
271. Tron, E.J., *The Optical Elements of the refractive error.*, in *Modern Trends in Ophthalmology*, F. Ridley and A. Sorsby, Editors. 1940, Paul B. Hoeber: New York.
272. Saunders, K.J., *Early refractive development in humans.* Surv Ophthalmol, 1995. **40**(3): p. 207-16.
273. Grosvenor, T., et al., *Houston Myopia Control Study: a randomized clinical trial. Part II. Final report by the patient care team.* Am J Optom Physiol Opt, 1987. **64**(7): p. 482-98.
274. Jensen, H., *Myopia progression in young school children and intraocular pressure.* Doc Ophthalmol, 1992. **82**(3): p. 249-55.
275. Chen, W.M. and G.R. Abecasis, *Family-based association tests for genomewide association scans.* Am J Hum Genet, 2007. **81**(5): p. 913-26.
276. Horvath, S. and N.M. Laird, *A discordant-sibship test for disequilibrium and linkage: no need for parental data.* Am J Hum Genet, 1998. **63**(6): p. 1886-97.
277. Clayton, D., *A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission.* Am J Hum Genet, 1999. **65**(4): p. 1170-7.
278. Weinberg, C.R., A.J. Wilcox, and R.T. Lie, *A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting.* Am J Hum Genet, 1998. **62**(4): p. 969-78.
279. Colhoun, H.M., P.M. McKeigue, and G. Davey Smith, *Problems of reporting genetic associations with complex outcomes.* Lancet, 2003. **361**(9360): p. 865-72.
280. van den Oord, E.J., *Controlling false discoveries in genetic studies.* Am J Med Genet B Neuropsychiatr Genet, 2008. **147B**(5): p. 637-44.
281. Sidak, Z., *On multivariate normal probabilities of rectangles: their dependence on correlations.* Ann Math Statist, 1968(39): p. 1425-1434.
282. Holm, S., *A simple sequentially rejective multiple test procedure.* Scand J Stat, 1979(6): p. 65-70.
283. Tsai, C.A., H.M. Hsueh, and J.J. Chen, *Estimation of false discovery rates in multiple testing: application to gene microarray data.* Biometrics, 2003. **59**(4): p. 1071-81.
284. Hsueh, H.M., J.J. Chen, and R.L. Kodell, *Comparison of methods for estimating the number of true null hypotheses in multiplicity testing.* J Biopharm Stat, 2003. **13**(4): p. 675-89.
285. Meuwissen, T.H. and M.E. Goddard, *Bootstrapping of gene-expression data improves and controls the false discovery rate of differentially expressed genes.* Genet Sel Evol, 2004. **36**(2): p. 191-205.
286. Garcia-Closas, M., et al., *Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash.* Cancer Epidemiol Biomarkers Prev, 2001. **10**(6): p. 687-96.
287. Harty, L.C., et al., *Collection of buccal cell DNA using treated cards.* Cancer Epidemiol Biomarkers Prev, 2000. **9**(5): p. 501-6.
288. Freeman, B., et al., *DNA by mail: an inexpensive and noninvasive method for collecting DNA*

samples from widely dispersed populations. Behav Genet, 1997. **27**(3): p. 251-7.

289. Le Marchand, L., et al., *Feasibility of collecting buccal cell DNA by mail in a cohort study.* Cancer Epidemiol Biomarkers Prev, 2001. **10**(6): p. 701-3.

290. Haque, K.A., et al., *Performance of high-throughput DNA quantification methods.* BMC Biotechnol, 2003. **3**: p. 20.

291. Duewer, D.L., et al., *NIST mixed stain studies #1 and #2: interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples.* J Forensic Sci, 2001. **46**(5): p. 1199-210.

292. Sham, P., et al., *DNA Pooling: a tool for large-scale association studies.* Nat Rev Genet, 2002. **3**(11): p. 862-71.

293. Garcia-Closas, M., et al., *Quantitation of DNA in buccal cell samples collected in epidemiological studies.* Biomarkers, 2006. **11**(5): p. 472-9.

294. Nielsen, K., et al., *Comparison of five DNA quantification methods.* Forensic Sci Int Genet, 2008. **2**(3): p. 226-30.

295. Lench, N., P. Stanier, and R. Williamson, *Simple non-invasive method to obtain DNA for gene analysis.* Lancet, 1988. **1**(8599): p. 1356-8.

296. Lum, A. and L. Le Marchand, *A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies.* Cancer Epidemiol Biomarkers Prev, 1998. 7(8): p. 719-24.

297. Walker, A.H., et al., *Collection of genomic DNA by buccal swabs for polymerase chain reaction-based biomarker assays.* Environ Health Perspect, 1999. **107**(7): p. 517-20.

298. Heath, E.M., et al., *Use of buccal cells collected in mouthwash as a source of DNA for clinical testing.* Arch Pathol Lab Med, 2001. **125**(1): p. 127-33.

299. King, I.B., et al., *Buccal cell DNA yield, quality, and collection costs: comparison of methods for large-scale studies.* Cancer Epidemiol Biomarkers Prev, 2002. **11**(10 Pt 1): p. 1130-3.

300. Bergen, A.W., et al., *Comparison of yield and genotyping performance of multiple displacement amplification and OmniPlex whole genome amplified DNA generated from multiple DNA sources.* Hum Mutat, 2005. **26**(3): p. 262-70.

301. Fan, J.B., et al., *Illumina universal bead arrays.* Methods Enzymol, 2006. **410**: p. 57-73.

302. Aas, J.A., et al., *Defining the normal bacterial flora of the oral cavity.* J Clin Microbiol, 2005. **43**(11): p. 5721-32.

303. Paster, B.J., et al., *Bacterial diversity in human subgingival plaque.* J Bacteriol, 2001. **183**(12): p. 3770-83.

304. Haffajee, A.D. and S.S. Socransky, *Relationship of cigarette smoking to the subgingival microbiota.* J Clin Periodontol, 2001. **28**(5): p. 377-88.

305. Shiloah, J., M.R. Patters, and M.B. Waring, *The prevalence of pathogenic periodontal microflora in healthy young adult smokers.* J Periodontol, 2000. **71**(4): p. 562-7.

306. Konig, K.G., *Diet and oral health.* Int Dent J, 2000. **50**(3): p. 162-74.

307. Glei, M., et al., *Assessment of DNA damage and its modulation by dietary and genetic factors in*

smokers using the Comet assay: a biomarker model. Biomarkers, 2005. **10**(2-3): p. 203-17.

308. D'Souza, G., et al., *Analysis of the effect of DNA purification on detection of human papillomavirus in oral rinse samples by PCR.* J Clin Microbiol, 2005. **43**(11): p. 5526-35.

309. Rudney, J.D. and R. Chen, *The vital status of human buccal epithelial cells and the bacteria associated with them.* Arch Oral Biol, 2006. **51**(4): p. 291-8.

310. Wilson, I.G., *Inhibition and facilitation of nucleic acid amplification.* Appl Environ Microbiol, 1997. **63**(10): p. 3741-51.

311. Cler, L., et al., *A comparison of five methods for extracting DNA from paucicellular clinical samples.* Mol Cell Probes, 2006. **20**(3-4): p. 191-6.

312. Fingert, J.H., et al., *Myocilin glaucoma.* Survey of Ophthalmology, 2002. **47**(6): p. 547-561.

313. Stone, E.M., et al., *Identification of a gene that causes primary open angle glaucoma.* Science, 1997. **275**(5300): p. 668-670.

314. Tamm, E.R., *Myocilin and glaucoma: facts and ideas.* Progress in Retinal and Eye Research, 2002. **21**(4): p. 395-428.

315. Polansky, J.R., et al., *Cellular pharmacology and molecular biology of the trabecular meshwork inducible glucocorticoid response gene product.* Ophthalmologica, 1997. **211**(3): p. 126-139.

316. Polansky, J.R., D.J. Fauss, and C.C. Zimmerman, *Regulation of TIGR/MYOC gene expression in human trabecular meshwork cells.* Eye, 2000. **14**: p. 503-514.

317. Tamm, E.R., et al., *Modulation of myocilin/TIGR expression in human trabecular meshwork.* Investigative Ophthalmology & Visual Science, 1999. **40**(11): p. 2577-2582.

318. Wang, L., et al., *Pro370Leu mutant myocilin disturbs the endoplasm reticulum stress response and mitochondrial membrane potential in human trabecular meshwork cells.* Molecular Vision, 2007. **13**(65-67): p. 618-625.

319. Wu, H., X.H. Yu, and E.P. Yap, *Allelic association between trabecular meshwork-induced glucocorticoid response (TIGR) gene and severe myopia.* Investigative Ophthalmology and Visual Science, 1999. **40**(4): p. S600.

320. Xu, L., et al., *High myopia and glaucoma susceptibility: The Beijing Eye Study.* Ophthalmology, 2006. **114**: p. 216-220.

321. Mitchell, P., et al., *The relationship between glaucoma and myopia: the Blue Mountains Eye Study.* Ophthalmology, 1999. **106**(10): p. 2010-5.

322. Wong, T.Y., et al., *Refractive errors, intraocular pressure, and glaucoma in a white population.* Ophthalmology, 2003. **110**(1): p. 211-217.

323. Young, F.A., *The nature and control of myopia.* J Am Optom Assoc, 1977. **48**(4): p. 451-7.

324. Junghans, B.M., et al., *A role for choroidal lymphatics during recovery from form deprivation myopia?* Optom Vis Sci, 1999. **76**(11): p. 796-803.

325. Leydolt, C., O. Findl, and W. Drexler, *Effects of change in intraocular pressure on axial eye length and lens position.* Eye, 2008. **22**(5): p. 657-61.

326. Schmid, K., *Myopia Manual.* 2004: Pagefree Publishing.

327. Nickla, D.L., C. Wildsoet, and J. Wallman, *The circadian rhythm in intraocular pressure and its relation to diurnal ocular growth changes in chicks.* Exp Eye Res, 1998. **66**(2): p. 183-93.

328. Tokoro, T., M. Funata, and Y. Akazawa, *Influence of intraocular pressure on axial elongation.* J Ocul Pharmacol, 1990. **6**(4): p. 285-91.

329. Quinn, G.E., et al., *Association of intraocular pressure and myopia in children.* Ophthalmology, 1995. **102**(2): p. 180-5.

330. Nomura, H., et al., *The relationship between intraocular pressure and refractive error adjusting for age and central corneal thickness.* Ophthalmic Physiol Opt, 2004. **24**(1): p. 41-5.

331. Tiburtius, H. and K. Tiburtius, *[New treatment possibilities of progressive school myopia].* Klin Monatsbl Augenheilkd, 1991. **199**(2): p. 120-1.

332. Goldschmidt, E., *Myopia in humans: can progression be arrested?* Ciba Found Symp, 1990. **155**: p. 222-9; discussion 230-4.

333. Edwards, M.H. and B. Brown, *IOP in myopic children: the relationship between increases in IOP and the development of myopia.* Ophthalmic Physiol Opt, 1996. **16**(3): p. 243-6.

334. Lee, A.J., et al., *Intraocular pressure associations with refractive error and axial length in children.* Br J Ophthalmol, 2004. **88**(1): p. 5-7.

335. Manny, R.E., et al., *IOP, myopic progression and axial length in a COMET subgroup.* Optom Vis Sci, 2008. **85**(2): p. 97-105.

336. Rohrer, B., J. Tao, and W.K. Stell, *Basic fibroblast growth factor, its high- and low-affinity receptors, and their relationship to form-deprivation myopia in the chick.* Neuroscience, 1997. **79**: p. 775-787.

337. Leung, Y.F., et al., *TIGR/MYOC proximal promoter GT-repeat polymorphism is not associated with myopia.* Human Mutation, 2000. **16**(6): p. 533.

338. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-265.

339. Farbrother, J.E., et al., *Linkage analysis of the genetic loci for high myopia on chromosomes 18p, 12q and 17q in 51 UK families.* Investigative Ophthalmology and Visual Science, 2004. **45**: p. 2879-2885.

340. Wigginton, J.E. and G.R. Abecasis, *PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data.* Bioinformatics, 2005. **21**(16): p. 3445-3447.

341. Vitale, S., et al., *Prevalence of refractive error in the United States, 1999-2004.* Archives of Ophthalmology, 2008. **126**(8): p. 1111-1119.

342. Katz, J., J.M. Tielsch, and A. Sommer, *Prevalence and risk factors for refractive errors in an adult inner city population.* Investigative Ophthalmology and Visual Science, 1997. **38**(2): p. 334-340.

343. Epstein, M.P., et al., *Genetic association analysis using data from triads and unrelated subjects.* Am J Hum Genet, 2005. **76**(4): p. 592-608.

344. Knapp, M., *A note on power approximations for the transmission/disequilibrium test.* American Journal of Human Genetics, 1999. **64**(4): p. 1177-1185.

345. Wentz-Hunter, K., X. Shen, and B. Yue, *Distribution of myocilin, a glaucoma gene product, in human corneal fibroblasts.* Molecular Vision, 2003. **9**(42-43): p. 308-314.

346. Nicodemus, K.K., A. Luna, and Y.Y. Shugart, *An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium-based statistical methods under different family structures, missing parental data, and population stratification.* Am J Hum Genet, 2007. **80**(1): p. 178-85.

347. Nakanishi, H., et al., *Absence of association between COL1A1 polymorphisms and high myopia in the Japanese population.* Invest Ophthalmol Vis Sci, 2009. **50**(2): p. 544-50.

348. Barrett, J.C. and L.R. Cardon, *Evaluating coverage of genome-wide association studies.* Nat Genet, 2006. **38**(6): p. 659-62.

349. Clark, A.G. and J. Li, *Conjuring SNPs to detect associations.* Nat Genet, 2007. **39**(7): p. 815-6.

350. Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. **69**(1): p. 124-37.

351. Thomson, G., *Mapping disease genes: family-based association studies.* Am J Hum Genet, 1995. **57**(2): p. 487-98.

352. Hintsanen, P., et al., *An empirical comparison of case-control and trio based study designs in high throughput association mapping.* J Med Genet, 2006. **43**(7): p. 617-24.

353. Royce, P.M. and B. Steinmann, *Connective Tissue and Its Heritable Disorders: Molecular, Genetic and Medical Aspects*. Second Edition ed. 2002: WileyBlackwell.

354. McBrien, N.A. and A. Gentle, *Role of the sclera in the development and pathological complications of myopia.* Prog Retin Eye Res, 2003. **22**(3): p. 307-38.

355. Rada, J.A., S. Shelton, and T.T. Norton, *The sclera and myopia.* Exp Eye Res, 2006. **82**(2): p. 185-200.

356. Serini, G. and G. Gabbiani, *Mechanisms of myofibroblast activity and phenotypic modulation.* Exp Cell Res, 1999. **250**(2): p. 273-83.

357. Gabbiani, G., *The myofibroblast in wound healing and fibrocontractive diseases.* J Pathol, 2003. **200**(4): p. 500-3.

358. Curtin, B.J. and C.C. Teng, *Scleral changes in pathological myopia.* Trans Am Acad Ophthalmol Otolaryngol, 1958. **62**(6): p. 777-88; discussion 788-90.

359. Funata, M. and T. Tokoro, *Scleral change in experimentally myopic monkeys.* Graefes Arch Clin Exp Ophthalmol, 1990. **228**(2): p. 174-9.

360. McBrien, N.A., L.M. Cornell, and A. Gentle, *Structural and ultrastructural changes to the sclera in a mammalian model of high myopia.* Invest Ophthalmol Vis Sci, 2001. **42**(10): p. 2179-87.

361. Gentle, A., et al., *Collagen gene expression and the altered accumulation of scleral collagen during the development of high myopia.* J Biol Chem, 2003. **278**(19): p. 16587-94.

362. Austin, B.A., et al., *Altered collagen fibril formation in the sclera of lumican-deficient mice.* Invest Ophthalmol Vis Sci, 2002. **43**(6): p. 1695-701.

363. McBrien, N.A., et al., *Expression of collagen-binding integrin receptors in the mammalian sclera*

and their regulation during the development of myopia. Invest Ophthalmol Vis Sci, 2006. **47**(11): p. 4674-82.

364. Bailey, A.J., *Structure, function and ageing of the collagens of the eye.* Eye, 1987. **1 (Pt 2)**: p. 175-83.

365. Siegwart, J.T., Jr. and T.T. Norton, *The time course of changes in mRNA levels in tree shrew sclera during induced myopia and recovery.* Invest Ophthalmol Vis Sci, 2002. **43**(7): p. 2067-75.

366. Dalgleish, R., *The human type I collagen mutation database.* Nucleic Acids Res, 1997. **25**(1): p. 181-7.

367. Larsen, J.S., *The sagittal growth of the eye. 3. Ultrasonic measurement of the posterior segment (axial length of the vitreous) from birth to puberty.* Acta Ophthalmol (Copenh), 1971. **49**(3): p. 441-53.

368. Suri, S. and R. Banerjee, *Biophysical Evaluation of Vitreous Humor, its Constituents and Substitutes.* Trends Biomater. Artif. Organs, 2006. **20**(1): p. 72-77.

369. Bishop, P.N., *Structural macromolecules and supramolecular organisation of the vitreous gel.* Prog Retin Eye Res, 2000. **19**(3): p. 323-44.

370. Morita, H., M. Funata, and T. Tokoro, *A clinical study of the development of posterior vitreous detachment in high myopia.* Retina, 1995. **15**(2): p. 117-24.

371. Liberfarb, R.M., et al., *The Stickler syndrome: genotype/phenotype correlation in 10 families with Stickler syndrome resulting from seven mutations in the type II collagen gene locus COL2A1.* Genet Med, 2003. **5**(1): p. 21-7.

372. Martin, E.R., et al., *A test for linkage and association in general pedigrees: the pedigree disequilibrium test.* Am J Hum Genet, 2000. **67**(1): p. 146-54.

373. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.

374. Santure, A.W. and H.G. Spencer, *Influence of mom and dad: quantitative genetic models for maternal effects and genomic imprinting.* Genetics, 2006. **173**(4): p. 2297-316.

375. Wojciechowski, R., et al., *Heritability of refractive error and familial aggregation of myopia in an elderly American population.* Invest Ophthalmol Vis Sci, 2005. **46**(5): p. 1588-92.

376. Chen, C.Y., et al., *Heritability and shared environment estimates for myopia and associated ocular biometric traits: the Genes in Myopia (GEM) family study.* Hum Genet, 2007. **121**(3-4): p. 511-20.

377. Guggenheim, J.A., et al., *Correlations in refractive errors between siblings in the Singapore Cohort Study of Risk factors for Myopia.* Br J Ophthalmol, 2007. **91**(6): p. 781-4.

378. Krause, U.H., et al., *The development of myopia up to the age of twenty and a comparison of refraction in parents and children.* Arctic Med Res, 1993. **52**(4): p. 161-5.

379. Shi, M., D.M. Umbach, and C.R. Weinberg, *Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families.* Am J Hum Genet, 2007. **81**(1): p. 53-66.

380. Wilcox, A.J., C.R. Weinberg, and R.T. Lie, *Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads".* Am J Epidemiol, 1998. **148**(9): p. 893-901.

APPENDIX 1.

THE INFORMATION PACK FOR PROVISIONAL PARTICIPANTS OF THE FAMILY STUDY OF MYOPIA

The information pack sent out to the provisional participants of The Family Study of Myopia consisted of the following:

1. Information sheet about the research project

Information about the research project

We would like to invite you to take part in The Family Study of Myopia, a research project investigating the genetic factors that lead to the development of high myopia (also known as short-sightedness).

What is the purpose of the study?

The study is investigating how myopia is inherited from one generation to the next. Our aim is to discover the genes that make some people more likely to become short-sighted than others. This will help our understanding of why myopia occurs, and in the future may aid the development of treatments for the condition.

Why have I been chosen?

We are seeking the participation of families from across the U.K. and Ireland in which there are one or more individuals with high myopia. We are looking for the help of about 200 such families in total.

Who is organising the study?

The study is organised by researchers from the Department of Optometry and Vision Sciences at Cardiff University and the Medical Genetics Department at the University of Wales College of Medicine. The research is funded by two eye research charities, the National Eye Research Centre and the College of Optometrists.

What would it involve if I take part?

- We would ask you to fill in a short questionnaire about your eyesight and your general health, and also to identify other members of your family who might be prepared to take part in the study (the more members of your family who are willing to take part in the study the better, even if these relatives are not short sighted themselves).

- To enable us to trace myopia genes in your family, we would ask you to provide two mouthwash samples. These mouthwashes are easily done by swishing some saline around in your mouth for 30 seconds. The equipment and instructions will be posted to you if you agree to take part. We can assure you that these samples will only be used for studying myopia genes, and that all samples will be coded in order to protect your anonymity

- We would ask for your permission to contact your Optometrist/Optician for details of your spectacle or contact lens prescription and your ocular health.

Will my confidentiality be maintained?
We take great care to ensure that the confidentiality of participating families is maintained. All personal details are kept securely, and the findings from this research will not identify individuals.

How do I participate?
If you would like to take part, please fill in the enclosed questionnaire and consent form and return them to us in the Freepost envelope provided. We will contact you with details about the mouthwash samples at a later date.

Contact for further information
If you have any further questions then we would be very happy to answer them either by telephone on 029 20875063, by post at the address overleaf or via email at *myopia@cardiff.ac.uk*.

Many thanks,

The Family Study of Myopia

2. Consent Form

Consent Form for the Family Study of Myopia *Please tick boxes*

I agree that my Optometrist/Optician can be contacted for further details about my eyes and health. ☐

I agree that other members of my family may be asked to take part in this study. ☐

I agree to provide mouthwash samples, which will be used to trace the passage of myopia genes through my family. ☐

I have been given an information sheet and have been given an opportunity to discuss the research. ☐

I understand that my participation is voluntary and that I am free to withdraw at any time without my legal rights being affected. ☐

I agree to take part in this study. ☐

Name	Date	Signature
Name of parent/guardian (if applicable)	Date	Signature
Researcher's name	Date	Signature

3. Study Questionnaire

Study Questionnaire

Title ☐ Mr. ☐ Mrs. ☐ Ms. ☐ Miss ☐ Other (Please specify)

Surname .....

First names .....

Date of birth .....

Address .....

.....

.....

Tel. Number .....

Please tick the box which you feel best describes your ethnic group:

- ☐ White European
- ☐ Other European
- ☐ African
- ☐ Asian
- ☐ American
- ☐ Afro-Caribbean
- ☐ Australasian
- ☐ Other (please specify)

1. At what age did you begin to wear spectacles?

   .....Yrs

2. Did you have any eye condition or eye disease at birth or in childhood?

   Yes ☐ No ☐ Don't know ☐

   If Yes, please give details.....

3. Do you currently suffer from any eye condition or disease?

   Yes ☐ No ☐ Don't know ☐

   If Yes, please give details.....

4. Have you had any eye surgery?

   Yes ☐ No ☐ Don't know ☐

   If Yes, please give details.....

5. Do you take any medication for your eyes?

   Yes ☐ No ☐ Don't know ☐

   If Yes, please give details.....

6. Were you born prematurely?

   Yes ☐ No ☐ Don't know ☐

7. Do you take any medication for any other health condition?

   Yes ☐ No ☐ Don't know ☐

   If Yes, please give

details.....

Please supply the names and addresses of any relatives who you think might be willing to participate in the research project. The participation of relatives who are not short-sighted is just as valuable as those that are. Similarly, the participation of your spouse would be very helpful, if appropriate.

Title ☐ Mr. ☐ Mrs. ☐ Ms. ☐ Miss ☐ Other (Please specify)

Surname .....
First names .....
Date of birth .....
Address .....
.....
.....
Tel. number .....

Is this relative short-sighted? ☐ Yes ☐ No ☐ Don't know

Please indicate their relationship to you:

☐ Sister ☐ Brother ☐ Mother ☐ Father ☐ Daughter ☐ Son
☐ Uncle ☐ Aunt ☐ Husband ☐ Wife ☐ Grandfather
☐ Grandmother

☐ Other (please specify): .....

****

Many thanks for your help

APPENDIX 2.

THE PROTOCOL FOR DNA EXTRACTION FROM MOUTHWASHES

PROTOCOL FOR DNA EXRACTION – PROTEINASE K FOR BUCCAL CELLS

Proteinase K solution: 1 x proteinase K buffer
10 mM Tris-HCL, pH 8.0
1 mM EDTA
0.5 % SDS
0.5 mg/ml proteinase K

1. Refrigerate mouthwashes (~15 ml) for at least 30 min, then centrifuge at 3500 rpm for 5 min.

2. Pour off the supernatant - ensuring you do not lose the pellet of buccal cells.

3. Add 380 µl 1 x Proteinase K buffer solution using a filter pipette tip. Pipette up-and-down to resuspend the cells, and transfer to a labelled 1.5 ml screw-cap vial. Freeze at -20°C until ready to process further.

4. Remove samples from freezer, thaw at 37°C, mix and spin.

5. Add 20 µl Proteinase K 10 mg/ml to each tube and incubate at 37°C for 2 hours, in a waterbath with continuous shaking (~100 rpm).

6. Centrifuge at 14000 rpm x 3 min to pellet insoluble material and transfer supernatant to a 1.5 ml silicon grease Eppendorf tube (use a syringe to instil ~100 µl silicon grease into a 1.5 Eppendorf tube, just underneath the hinge; centrifuge at 3000 rpm for 4 seconds with the hinge pointing outwards, to create a smear of grease down the side of the tube).

7. Add 470 µl of phenol/chlorophorm to the sample (phenol:chlorophorm:isoamyl alcohol – 25:24:1) and vortex vigorously for 30 seconds. Then centrifuge at 14000 rpm for 2 mintues.

9. If debris remains in the supernatant, transfer it to the second 1.5 ml silicon grease Eppendorf tube and repeat the phenol-chlorophorm extraction (steps 7-8).

10. When no debris remains in the supernatant, transfer it to a 1.5 ml screw-cap vial and add 19 µl of 5M NaCl. Mix and spin.

11. Add 1 ml of 100 % ethanol. Mix and spin. Then leave to precipitate at -20°C overnight.

13. Remove samples from freezer, invert a few times to mix, and then centrifuge at 14000 rpm for 10 minutes.

14. Discard the supernatant and add 1 ml of ice cold 70 % ethanol. Centrifuge at 14000 rpm for 2 minutes.

15. Remove the majority of supernatant , then use a narrow pipette tip to remove the last traces of ethanol, taking care not to lose the pellet.

16. Air-dry the tube in an inverted position for 3 minutes.

17. Resuspend the pellet in 50 µl of TE and incubate for 15 minutes at 37°C with periodic gentle vortexing to ensure that the pellet is fully dissolved.

LIST OF PUBLICATIONS

Publications arising from this work:

1. Quality of DNA extracted from mouthwashes.
Zayats T, Young TL, Mackey DA, Malecaze F, Calvas P, Guggenheim JA.
PLoS One. 2009 Jul 7;4(7):e6165.

2. COL1A1 and COL2A1 genes and myopia susceptibility: evidence of association and suggestive linkage to the COL2A1 locus.
Metlapally R, Li YJ, Tran-Viet KN, Abbott D, Czaja GR, Malecaze F, Calvas P, Mackey D, Rosenberg T, Paget S, Zayats T, Owen MJ, Guggenheim JA, Young TL.
Invest Ophthalmol Vis Sci. 2009 Sep;50(9):4080-6. Epub 2009 Apr 22.

3. Myocilin polymorphisms and high myopia in subjects of European origin.
Zayats T, Yanovitch T, Creer RC, McMahon G, Li YJ, Young TL, Guggenheim JA.
Mol Vis. 2009;15:213-22. Epub 2009 Jan 26.

3. Comment on 'A PAX6 gene polymorphism is associated with genetic predisposition to extreme myopia'.
Zayats T, Guggenheim JA, Hammond CJ, Young TL.
Eye. 2008 Apr;22(4):598-9; author reply 599. Epub 2008 Jan 25

Other publications produced during the PhD period:

1. Season of birth, daylight hours at birth, and high myopia.
McMahon G, Zayats T, Chen YP, Prashar A, Williams C, Guggenheim JA.
Ophthalmology. 2009 Mar;116(3):468-73. Epub 2009 Jan 20.

2. Axes of astigmatism in fellow eyes show mirror rather than direct symmetry.
Guggenheim JA, Zayats T, Prashar A, To CH.
Ophthalmic Physiol Opt. 2008 Jul;28(4):327-33.