The role of short tandem repeat genetic variation in myopia and other ocular traits

October 2022

A thesis submitted to Cardiff University for the degree of

Master of Philosophy

By

Jiangtian Cui

School of Optometry and Vision Sciences

Cardiff University

Supervised by

Prof. Jeremy A. Guggenheim

Dr. Louise Terry

AKNOWLEDGEMENTS

I would like to express my gratitude to Professor Jeremy A. Guggenheim for his supervision throughout this project. His huge support, patience and encouragement make this thesis possible. I would like to thank my co-supervisor Dr. Louise Terry and academic advisor Dr. Tony Redmond for their feedback on my work.

It has been my great privilege and pleasure to be able to work with Dr. Rosie Clark.

Thank you to all in the School of Optometry and Vision Science for your support throughout my MPhil.

I owe a great gratitude for my family for their massive support during my MPhil.

SUMMARY

The purpose of the study was to identify new genetic risk factors for the ophthalmic traits, strabismus and myopia. Short tandem repeats (STRs) are regions of the genome that contain repetitive sequences of DNA, e.g. CACACACA or ATTGATTGATTG. Their repetitive nature makes STRs prone to mutation during meiosis. Two hypotheses were examined by performing genetic association studies for STR markers in large samples of participants from the UK Biobank project.

An initial study was carried out examining the trait 'self-reported strabismus'. Previous research identified a single nucleotide polymorphism (SNP) on chromosome 17 that was strongly associated with strabismus. Therefore, tests for association of self-reported strabismus and STR markers on chromosome 17 were performed. The STR variant most strongly associated with the phenotype was Human_STR_613083. However, no marker retained evidence of association after accounting for multiple testing using the Bonferroni method.

In a second study, a case-control GWAS was performed in which cases with high myopia were compared to controls with moderate-to-high hyperopia. This GWAS identified two STRs associated with case-control status: Human_STR_827099 and Human_STR_424816, nearby the genes *PRSS56* and *SIX6* on chromosomes 2 and 14, respectively. Conditional analyses revealed the two lead STRs were in linkage disequilibrium with previously identified myopia-associated SNPs. As part of the second study, an analysis of STR genotypes on the X chromosome of male participants revealed that the error rate of STR genotyping was related to the STR motif length.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
avMSE	Average Mean Spherical Equivalent
bp	Basepair
C	Cytosine
CE	Capillary Electrophoresis
CI	Confidence Interval
CNV	Choroidal Neovascularization
CRAM	Compressed Reference-Oriented Alignment Map
CREAM	Consortium For Refractive Error And Myopia
D	Diopter
DNA	Deoxyribonucleic Acid
DZ	Dizygotic
EM	Expectation–Maximization
FDR	False Discovery Rate
FE	Functional Equivalence
FP	False Positive
FWER	Family-Wise Type 1 Error Rate
GRCh38	Genome Reference Consortium Human Build 38
GWAS	Genome-Wide Association Study
H^2	Broad-Sense Heritability
h^2	Narrow-Sense Heritability
HES	Hospital Episode Statistics
IQR	Interquartile Range
LD	Linkage Disequilibrium
logMAR	Logarithm of the Minimum Angle tf Resolution
MMD	Myopic Macular Degeneration
MR	Mendelian Randomization
MRI	Magnetic Resonance Imaging
MZ	Monozygotic
NGS	Next-Generation Sequencing
OPMD	Oculopharyngeal Muscular Dystrophy
OQFE	Original Quality Functional Equivalent
OR	Odds Ratio
ORF	Open Reading Frame
PC	Principal Components
PCR	Polymerase Chain Reaction
pVCF	Multi-Sample VCF
QC	Quality Control
Q-Q	Quantile-Quantile
QTL	Quantitative Trait Loci

RD	Retinal Detachment
RNA	Ribonucleic Acid
SCORM	Singapore Cohort of the Risk Factors for Myopia
SD-OCT	Spectral Domain Optical Coherence Tomography
SE	Spherical Equivalent
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
SV	Structural Variation
TP	True Positive
UK	United Kingdom
VA	Additive Genetic Effect
VCF	Virtual Contact File
VD	Non-Additive Genetic Effects
VG	Variation In Genetics
VP	Phenotypic Variation
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization

Chapter 1. General Introduction

1.1. Genetics

1.1.1. Monogenic and polygenic inheritance

Monogenic inheritance refers to the form of inheritance where the mutation of a single gene or allele could solely determine the trait. The inheritance pattern of single gene diseases is referred to as Mendelian, named after Gregor Mendel, the Augustinian monk and botanist who formulated the laws of heredity through careful breeding experiments in pea plants. Mendel first statistically summarized the different patterns for the selected traits in pea plants across generations, and recognized that gene segregation complies with the laws, presently known as 'Mendel's law of inheritance', determining the probability of recurrence of traits for subsequent generations. Monogenic inheritance of disease in a family can be established by observing the pattern of transmission, for which an accurate family history is important.

Most gene sequences occur in different versions in different individuals due to the accumulation of mutations; these different versions are referred to as 'alleles' and the sites at which they occur are called polymorphisms. Individuals carry two chromosomes. At a specific location on a chromosome, individuals can carry either a 'reference' (commonly occurring) allele or a 'alternate' (or 'rare') allele. The likelihood of carrying the alternate allele depends on the population frequency of the allele and its functional consequences. If the alternate allele greatly increases the risk of a phenotype or disease, the allele is sometimes called the disease-causing allele or disease-causing mutation. The expression of the mutated allele versus normal allele can be characterized as dominant, additive, or recessive. In total, there are five modes of inheritance for single-gene diseases: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive and mitochondrial. Different modes

of single-gene diseases refer to diseases inherited in different patterns depending on the genomic location of the gene and whether one or two alleles of the gene need to be functional for the pathology to manifest.

Polygenic inheritance, as opposed to monogenic inheritance, is characterized by traits that are controlled by two or more genes. The determinant genes are often very numerous in quantity but each with just a small effect on the phenotype. Polygenic gene polymorphisms typically have an additive effect on the phenotype. Mendelian inheritance patterns, the core notion of monogenic inheritance, are not applicable to analyze traits with polygenic determinism. Instead of just two options, polygenic traits generally have a continuous phenotypic spectrum. Examples of polygenic traits are skin colour, height and even intelligence.

Myopia, usually high myopia, sometimes appears as one of the features of a variety of rare heritable disease syndromes (Curtin 1985). Many of these rare syndromes, such as Stickler syndrome and Marfan syndrome, are single-gene disorders, while Down syndrome is due to an extra copy of chromosome 21. Accordingly, simple Mendelian inheritance patterns (dominant, recessive or sex linked) have been reported in some rare families with high myopia (Goss et al. 1988). By definition, such cases provide evidence that monogenic forms of myopia exist, which are caused by single-gene mutations.

More generally, evidence shows both heredity and environmental factors influence the development of myopia. Similar to many other ‘multifactorial’ diseases in nature, causes of myopia onset and development consist of both environmental and genetic factors. A ‘complex disease’ or ‘complex trait’ results from multiple genes, diverse environmental factors and potentially gene-gene or gene-environment interactions. Therefore, a single susceptibility gene is neither necessary not sufficient to cause a complex disease, and the underlying genetic effects involve probabilistic

predisposition (‘genetic susceptibility’) rather than predetermined programing (Tang et al. 2008). A complex disease or trait may segregate in families but will not show a typical Mendelian inheritance pattern. Type 2 diabetes shows characteristics typical of a complex disease, with rare monogenic forms showing typical Mendelian inheritance, while the more common cases are caused by a complex set of genetic and environmental risk factors.

1.1.2. Heritability: twin studies, family studies and population studies

It is often of interest to researchers to evaluate how much variation in a particular trait is the result of biological factors versus environmental factors. Heritability is the conception to quantify the genetic contribution to a trait – in other words, how much of the variation in a trait is controlled by genetic differences between individuals. The term heritability is often used to refer to the resemblance between parents and their offspring: a high heritability implies a robust resemblance between parents and offspring regarding a specific trait, while low heritability has the opposite implication.

Heritability is formally defined as the proportion of phenotypic variation (VP) that results from variation in genetics (VG). The ‘broad-sense’ heritability is defined as H^2 = VG/ VP. On the other hand, a finer model divides the variation in genetics into two classes, the additive genetic effect (VA) and non-additive genetic effects (VD), depending on the different types of inheritance. By definition, a formulation for the ‘narrow-sense’ heritability is h^2 = VA/ VP, in which only the proportion of genetic variation due to additive genetic variation is captured. Thus, there is a distinction between the two values, H^2 and h^2. In practice, h^2 can be a more useful measure for selection of animal or plant characteristics, because the additive genetic variation commonly responds to artificial selection. Moreover, the resemblance between relatives is largely driven by additive genetic variance.

Twins are a valuable resource for genetic studies in which the relative contribution of genetic background and environment is being investigated. Twins offer unique opportunities for studies beyond the analysis of phenotypic heritability (van Dongen et al. 2012). For example, comparison of discordant monozygotic (MZ) twins to search for disease-associated biological markers provides unique insight, since MZ twins are perfectly matched from a genetic perspective, apart from potential differences in DNA methylation patterns and accumulation of somatic mutations during a person's lifetime. Research aimed at testing the role of environmental conditions on the risk of a disease are usually susceptible to confounding, which hinders causal inference (Galton 1883). Twin studies provide an opportunity to find potential risk factors, since genetic variation is so well controlled. Twin research has become one of the favourite tools for behavioral geneticists and psychologists to estimate the heritability of traits and to quantify the effects either from shared environment within a family or individually unique environment on a specific trait.

Family-based genetic studies examining the inheritance patterns of discrete traits also have value (as applied in Mendel's early studies on pea plants). The inheritance pattern within pedigrees can often reveal whether a disease is monogenic or polygenic. Furthermore, traits that co-segregate with a specific genetic marker allow researchers to map, locate, and finally identify the relevant causative mutation. Compared with studies of unrelated individuals, studies of extended pedigrees or even nuclear families are especially well-suited for investigating the role of rare genetic variants that have large effect sizes.

Meanwhile, family studies are also valuable for analysis of polygenic traits. Family members share a predictable proportion of their genes 'identical-by-descent' (Borecki and Province 2008), which can be quantified as a function of the degree of their relationship (or 'kinship coefficient'). In the special case of MZ twins, who have

almost zero genetic variation among each other (and therefore a kinship coefficient = 1), trait variation is attributable to epigenetic phenomena and environmental factors. On the other hand, family members are likely to share more homogeneous environmental exposures, living closer together geographically, with similar socioeconomic status, and perhaps similar health-related living habits such as smoking, diet, alcohol intake, and habitual exercise. These similarities can reduce residual noise variance, thereby enhancing statistical power to detect relevant causal factors.

Finally, genetic studies of families also offer the technical advantage of being able to check for genotyping errors. A high rate of Mendelian inconsistencies, or markers that show significant deviations from Hardy-Weinberg equilibrium, can be signs of genotype error, sample mix-up, or other quality control problems (Borecki and Province 2008).

1.1.3. Overview of classes of genetic variation

Genetic variation is the difference in DNA sequences between individuals within a population, appearing in both germ cells and somatic cells. However, only the variation within germ cells is heritable and affects population dynamics, and ultimately evolution. Mutation and recombination are the two main sources of genetic variation.

Mutations are the original source of genetic variation, and the permanent alternation to DNA sequence. De novo (new) mutations occur when there is an error during DNA replication, which fails to get fixed by the cell's DNA repair machinery. If such an error gets copied in replication, then it becomes a mutation. Recombination is another major source of genetic variation. The recombination process occurs when DNA from

both parents is interchanged via the cutting and pasting (‘restriction’ and ‘ligation’) of chromosomal segments, when homologous DNA strands align and cross-over. Recombination creates a new combination of polymorphisms derived from each parent in the germ-cells.

Types of genetic variation are classified by the change in DNA sequence. These types include single base-pair substitutions, better known as ‘single nucleotide polymorphisms’ (SNPs), insertion or deletion polymorphisms (‘indels’), structural variation (SV), and short tandem repeats (STRs; also known as microsatellites). SNPs are variations in which a single DNA nucleotide base is substituted by another, e.g. an adenine (A) base may replace a cytosine (C). An indel refers to the insertion or deletion of a short stretch of DNA sequence that can range from 1 to hundreds of bases in length. SV describes the genetic variation that happens on a larger, consecutive fragment of DNA sequence. SVs include both copy number variation and chromosomal rearrangement events, such as insertion, deletion, inversion and duplication. An STR is a DNA motif unit 2-6 base pairs in length that is repeated multiple times. Often, the length of a specific STR in the genome is highly variable (polymorphic) within a population. Regions of the genome consisting of long stretches of repetitive DNA were once considered as “junk DNA”, as their role was difficult to define. Estimates suggest that approximate two thirds of the human genome is composed of DNA repeats (de Koning et al. 2011). Recent research has revealed the functional consequences of repeats include the (direct) generation of variability in gene expression, for example by altering transcription factor and enhancer binding (Meštrović et al. 2015) as well as the (indirect) regulation of gene expression via epigenetic modifications (Lannes et al. 2019). Common elements within repetitive segments include both transposable elements and short tandem repeats (STRs) (Biscotti et al. 2015), collectively representing a large portion of eukaryotic genomes (Charlesworth et al. 1994; López-Flores and Garrido-Ramos 2012). STRs comprise about 3% of the human genome (Lander et al. 2001). Further

information about STRs is presented in section 1.1.3.1.

1.1.3.1. Short tandem repeats: origin

Unlike SNPs, STR mutations usually lead to the gain or loss of an entire repeat motif, and sometimes even two or more repeats change together. Although the exact mechanism underlying such mutations is unclear, one proposed cause is the slippage of DNA during replication. A mismatch of the reference DNA strand and the other strand replicated during meiosis could cause a different number of copies of the repeat unit to be transmitted to germ cells (Tautz and Schlötterer 1994). DNA polymerase, the enzyme that reads and replicates DNA, can slip while moving along the DNA template strand and then re-start at a new, incorrect position. This form of replication mismatch is more likely to occur when a repetitive sequence is read, which means a higher error rate for STR replication compared to the replication rate elsewhere in the genome. Several studies have assessed the occurrence rate of slippage-caused STR mutation; currently, it is estimated to occur about once per 1,000 meioses (Klintschar et al. 2004; Forster et al. 2015). Thus, the STR mutation rate in repetitive DNA is three orders of magnitude more common than single-nucleotide changes.

1.1.3.2. Short tandem repeats: monogenic diseases associated with expansions

Certain STRs appear to be highly unstable, in that the number of repeats can expand to remarkably high levels. Currently, 40 inherited human disorders have been found to be association with grossly expanded STRs (Paulson 2018a). These disorders mostly affect the nervous system; they include Huntingdon's disease, fragile X syndrome, spinocerebellar ataxia, Friedreich's ataxia, and amyotrophic lateral

sclerosis/frontotemporal dementia (Paulson 2018a). Recently, Fuchs endothelial corneal dystrophy was found to be causal by expansion of a CTG repeat in the TCF4 gene (Fautsch et al. 2021). Disease-causing STRs can be located in the coding region of genes, the 5' or 3' untranslated regions, or in introns.

The high degree of variability in the length of repeat expansions has a major impact on disease severity. It also influences whether the disorder follows conventional Mendelian rules of inheritance. Traditionally, Mendelian inheritance assumes that mutations are fixed, and therefore that there is a clear-cut autosomal dominant, autosomal recessive, or X-linked pattern of inheritance within families and across generations. STR expanded repeats exhibit dynamic mutation of their length across generations, such that these disorders only loosely follow Mendelian rules of transmission.

Although STR expansion diseases manifest in a diversity of phenotypes, the underlying genetic patterns share general features. For instance, all disease-associated repeat expansions arise from polymorphic repeats present in the normal population. The length of the STR in unaffected individuals varies across a relatively narrow scale, which differs from disease to disease. In general, repeats at the high end of the normal range are at an increased risk of further (abnormal) expansion during transmission to the next generation, with the chance that the expansion length moves into the pathogenic range. Therefore, a person can inherit a repeat expansion disease even if they have no family history of the disease (so-called "sporadic cases").

1.1.3.3. Short tandem repeats: contribution to polygenic traits

As mentioned before, expansions at several dozen STRs have been found to cause

Mendelian disorders (Mirkin 2007) such as Huntingdon's Disease and hereditary ataxias. Limited by the power of bioinformatics analysis of repetitive regions, many STRs are often precluded from genome wide studies (Li 2014a). Thus, many STRs in the human genome, alongside the pathogenic STRs that cause monogenic inheritance diseases, are lacking of investigation. However, recent studies are increasingly reporting the association of STR polymorphisms with complex traits such as gene expression levels (Nasrallah et al. 2012; Gymrek et al. 2016; Quilez et al. 2016).

1.1.3.4. Short tandem repeats: mechanisms of action

STRs may influence gene expression levels through diverse mechanisms (Gemayel et al. 2010). STR expansions can be located in both open reading frames (ORFs) and promoter regions, through which the phenotypes could be mediated. Variations in STR repeat length in coding regions can directly affect protein function, e.g. via frameshifts that result in mistranslated, nonfunctional proteins. On the other hand, variations in STRs located within promoters can free or block RNA polymerase binding sites, resulting in either higher expression or reduced expression. For example, the CCG repeat linked with Fragile X Syndrome leads to the disruption of DNA methylation, which alters the expression of the *FMR1* gene (Liu et al. 2018). Dinucleotide repeats may change the affinity of nearby enhancer or repressor binding sites (Afek et al. 2014). Furthermore, certain STR repeats may alter the DNA or RNA strands into non-canonical secondary structures such as G-quadruplexes (Conlon et al. 2016), R-loops (Lin et al. 2010), or Z-DNA (Rothenburg et al. 2001), which can also alter transcriptional activity.

STR expansion diseases are often characterized by a genotype-phenotype correlation between the repeat length and the severity of signs and symptoms. Longer repeat

expansions are usually linked with a more severe disease phenotype and an earlier age-of-onset. For at least 10 diseases, including Huntingdon's disease, spinal and bulbar muscular atrophy, dentatorubral-pallidoluysian atrophy and seven Spinocerebellar ataxias, there is a significant correlation between the CAG motif repeat length and disease severity, along with a reverse correlation between the repeat length and the age-of-onset of symptoms. Very long repeats usually act as a causal variant for earlier disease and more severe signs and symptoms (Doyu et al. 1992; Igarashi et al. 1992; Andrew et al. 1993; Snell et al. 1993; Stine et al. 1993; Illarioshkin et al. 1994; Kawaguchi et al. 1994; Komure et al. 1995; Schöls et al. 1995; Penney et al. 1997; Johansson et al. 1998; Rosenblatt et al. 2012; Figueroa et al. 2017), contributing to 40% to 75% of the effects. The remaining variation in the age-of-onset is the result of other minor genetic modifiers and environmental factors; currently, these factors are poorly understood (Paulson 2018b).

The pathogenic mechanisms underlying repeat expansion diseases can be highly diverse depending on the specific mutation. For simplicity, the primary influence is often classified as a toxic gain-of-function or deleterious loss-of-function (Table 1.1). Relatively few diseases are found with a loss of function mechanism: these disorders are inherited in an autosomal recessive manner or an X-linked recessive manner. As a contrast, disorders with a toxic gain-of-function mechanism (also known as a dominant negative effect) are inherited in an autosomal dominant manner or X-linked dominant manner (Paulson 2018b).

Table 1.1 Primary mechanism of disease (reproduced by published study (Paulson, 2018b))
Toxic gain of function
CAG/polyQ diseases
Spinocerebellar ataxia types 8, 10, 12, 31, 36
C9ORF72 frontotemporal dementia/ amyotrophic lateral sclerosis
Huntington disease-like 2
Myotonic dystrophy types 1 and 2
Oculopharyngeal muscular dystrophy
Fragile X tremor ataxia syndrome
Fuchs endothelia corneal dystrophy
Loss of function
Friedreich Ataxia
Fragile X syndrome
Myoclonic epilepsy (Unverricht-Lundborg)

Approximately six different molecular mechanisms by which STR expansions cause disease have been discovered. Certain STRs alter the folding and thus the intramolecular structures of DNA or RNA, which can, in turn, affect either transcription, translation, or the binding of various RNA binding proteins (Chen et al. 2017). For example, the GAA repeat that causes Friedreich ataxia produces a DNA-RNA triplex structure that impedes transcription; this results in a marked reduction in the expression of the encoded protein, frataxin. For repeats located within the coding sequence of the exome, a trinucleotide repeat expansion can lead to an unusually long run of a specific amino acid. Such an STR expansion will have a direct influence on the disease. For example, the so-called 'polyglutamine diseases' – including Huntington's disease and several types of spinocerebellar ataxia – are caused by a repeated expansion of the CAG motif, which codes for the amino acid glutamine. Similarly, oculopharyngeal muscular dystrophy (OPMD) is caused by a pathogenic protein containing an expanded tract of alanine amino acids, encoded by a GCG repeat STR.

The length of STR expansions has a strong correlation with the disease severity and age-of-onset. Due to the high probability of slippage events during DNA replication, STRs exhibit high variability in their numbers of motif repeats across the population. The length range varies widely, for instance lying within the range 8-18 repeats in OPMD to more than a thousand repeats in myotonic dystrophy and spinocerebellar ataxia. Expansions residing in protein-coding regions usually have shorter lengths, probably because of the necessity to encode a functional protein. Any change in the STR length would potentially have a direct functional impact on the encoded protein, sufficient to create strong pressure for natural selection. In contrast, the STR expansions with the largest length are commonly found in introns or untranslated regions. These regions are more tolerant of the size of expansions and exhibit a less clear-cut correlation between repeat length and disease severity.

1.1.4. DNA sequencing and genotyping

1.1.4.1. Whole exome and whole genome sequencing

Whole genome sequencing (WGS) is the process of determining the entire DNA sequence of an organism's genome at a single time. For human beings, WGS is used to determine nearly all of the approximately 3 billion nucleotides of an individual's complete DNA sequence, including non-coding sequence. As of 2017, between 4% to 9% of the human genome had still not been sequenced, mostly due to technical difficulty of sequencing these highly repetitive, GC-rich regions. Recently, the Telomere-to-Telomere Consortium presented a complete 3.055 billion–base pair sequence of a human genome, *T2T-CHM13*, that included gapless assemblies for all chromosomes except chromosome Y (Nurk et al. 2022). Some of the largest reference gaps include human satellite, repeat arrays and the short arms of all five acrocentric chromosomes, which in GRCh38 are represented as multi-megabase stretches of unknown bases. Long-read resequencing studies are now needed to identify the polymorphic variation and reveal any potential phenotypic associations within these regions.

DNA sequencing techniques have advanced greatly in recent years. These advances now make WGS a routine procedure. In the 1970s and 1980s, manual sequence tools, such as Maxam-Gilbert sequencing and Sanger sequencing were initially used to sequence the whole bacteriophage and animal viral genomes. One could use individual Sanger sequencing reactions to cover any desired region, but this testing approach can be costly, making it suitable for analysis of just a small subset of genes or a gene-specific test.

Shotgun sequencing, a later technology for DNA sequencing, firstly successfully sequenced almost the entire human genome in 2000. In shotgun sequencing, DNA is broken up randomly into many small fragments and read by Sanger sequencing. Multiple overlapping fragments are sequenced. Subsequently, a computer program joins the fragments into a continuous sequence. Although the technology realized the first human genome sequencing, it was still too costly and time-consuming to be applied widely.

Since 2005, high-throughput sequencing (or next-generation sequencing, NGS) gradually replaced the former tools, due to its high speed and affordable price. These technologies use the concept of massively parallel processing, reading 1 million to 43 billion short reads (50-400 bases each) per instrument run (Mukhopadhyay 2009). New high-throughput sequencing technologies have contributed to a significant reduction in the cost for sequencing - nearing the mark of $1000 per genome (von Bubnoff 2008). NGS with either whole genome (WGS) or whole exome (WES) sequencing, is now a standard clinical test for many individuals with suspected genetic disorders.

The development of next-generation-sequencing has allowed WGS, WES and the analysis of specific candidate genes to be adopted by scientists and clinicians (Pottier et al. 2015; Williams et al. 2016; Bonvicini et al. 2019). As WGS has become more affordable, the analysis of STRs has been revolutionized. WGS-based STR detection tools, such as tandem repeat finder, can detect novel STRs from assembled genome sequences, including the human genome (Gelfand et al. 2007). New software tools and pipelines such as superSTR (Fearnley et al. 2022) can also be directly applied for STR profiling in WGS data, making these variants more accessible to study by researchers.

Long-read sequencing techniques such as PacBio and Oxford Nanopore sequencing

can generate long reads with up to hundreds of thousands of basepairs. Long reads can provide better alignment for pathogenic STRs by using information from flanking sequences and ensuring that the full-length STR is accessible for analysis. Several computation tools have been developed to determine repeat counts based on long-read sequence data, such as RepeatHMM (Liu et al. 2017), Tandem-genotypes (Mitsuhashi et al. 2019), RepLong (Guo et al. 2018), and TRiCoLOR (Bolognini et al. 2020). Historically, one drawback that has limited the adoption of long-read sequencing is the relatively high base-calling error rate (3 to 15%). However, new developments have largely overcome this limitation, and long-read sequencing is likely to become increasingly popular for genotyping STRs.

1.1.4.2. Array-based genotyping

Genotyping arrays, often called SNP arrays, are another efficient high-throughput DNA reading technology that originated from the early 2000s. SNP arrays are a powerful platform for simultaneously analyzing hundreds of thousands of SNPs. The SNP arrays contain immobilized allele-specific oligonucleotide probes, which are hybridized with fragmented nucleic acid sequences of target DNA, labelled with fluorescent dyes. The hybridization signals will ultimately be detected by an imaging system, and decoded to infer the SNP genotype information. In research, SNP arrays are most widely used for GWA studies, in which SNP-based genetic analysis can be used for disease associated loci mapping to determine the susceptibility genes in individuals.

1.1.4.3. Genotyping short tandem repeats

Genotyping STRs is challenging. The classic method for genotyping STRs is capillary

electrophoresis (CE). In the CE method, DNA molecules migrate through an electrolyte solution within a glass capillary under the influence of an electric field. DNA molecules are separated according to their ionic mobility. Larger DNA molecules move more slowly through the capillary than the smaller DNA fragments. DNA samples from patients are mixed with a set of purified DNA fragments of known size that serve as size reference markers. The electric fields required for CE are strong (e.g. 300 V/cm), which produces much fasters runs for CE compared to gel electrophoresis (Biscotti et al. 2015). The CE technique has very high accuracy but because of low throughput, it is limited to relatively small scale studies.

STRs can also be genotyped using data obtaining from standard next-generation DNA sequencing - either whole genome sequencing or whole exome sequencing. Until recently, this method of STR genotyping has lacked adequate tools (Treangen and Salzberg 2011). One major problem is that not all NGS reads that align to an STR are informative. For example, only if a single or paired-end sequencing read entirely encompasses an STR locus can the read be used for exact STR genotyping. While reads partially encompassing an STR do provide information about the minimum repeat length, the true length is unknown. Another problem is the difficulty of analyzing STRs containing insertions/deletions, while being tolerant of computer processing time (Li and Homer 2010). Finally, technical artefacts from the PCR amplification process introduce noise into the final length measurements of STRs. Because of the repetitive nature of STRs, PCR amplification causes 'stutter' noise. This artefact is due to successive slippage events of Taq DNA polymerase during amplification; some DNA copies will contain a different number of repeats compared to the original DNA template.

Five years ago, there was no efficient analysis method available for detecting and genotyping *novel* STRs. Since then, four new methods to detect repeat expansion have been introduced: ExpansionHunter (Dolzhenko et al. 2017), exSTRa (Tankard et

al. 2018), STRetch (Dashnow et al. 2018), and TREDPARSE (Tang et al. 2017). These repeat expansion detection methods all require paired-end sequencing data. Between the opposite read pairs, there is typically a stretch of DNA that is not sequenced. The key to repeat expansion detection is to assess the lengths of reads that partially or entirely encompass an STR.

There are also several software packages designed for genotyping STRs already known to exist in specific regions of the genome. These software packages enable high-throughput genotyping even in large cohorts, which would not be practical for capillary electrophoresis. These packages include: lobSTR (Gymrek et al. 2012), HipSTR (Willems et al. 2017) and RepeatSeq (Highnam et al. 2013). The HipSTR program is designed to genotype STRs from Illumina sequencing data. It utilises a highly flexible realignment framework, effectively mitigating stutter noise from PCR. The HipSTR program is designed to genotype population-scale data, which enables further GWAS studies of STRs in cohorts. HipSTR displayed good performance in a published whole genome sequencing data study (Willems et al. 2017).

The phasing procedure can help to identify alleles on maternal and paternal chromosomes. This information is often important for understanding gene expression patterns for genetic disease. As an example, for a patient with a recessive form of eye disease, if two mutations in the disease gene are identified, then the two mutations are likely to be the causal variants of the eye disease if they occur on the maternal and paternal chromosomes, respectively. However, the two mutations would be excluded as the sole causal variants if they are mapped to the same chromosome (i.e. the two mutations are in phase). Therefore, additional information about phasing can increase the accuracy of identification of causal variants.

The process of statistically estimating haplotypes from genotype data can be performed in HipSTR with population-based genotype data. However, differences in

linkage disequilibrium patterns in populations of differing ancestry limits the degree of accuracy for phasing. Also, with genotype data, the current methodologies are not able to reliably phase small segments under 5cMs. For population-based phasing, a false positive rate of over 67% for 2-4 cM segments was reported (Durand et al. 2014). In the future, datasets generated using long-read sequencing techniques, such as PacBio or Nanopore, could be a solution to the limitation of the phasing accuracy.

1.1.5. Genetic analysis

1.1.5.1. Identifying monogenic disease genes (linkage analysis and sequencing)

Linkage analysis is the classic method for mapping the genes for heritable traits to their chromosomal locations. Linkage analyses are conducted in families, within which heritable traits are segregated in pedigrees. The traits can be binary, having only two values, such as absence or presence of a disease, or quantitative (continuous) such as body mass index. A genetic marker that is located nearby in the genome to a mutation or polymorphism that affects a trait or disease will exhibit 'co-segregation' (co-inheritance of the phenotype and marker alleles amongst individuals in a pedigree). To carry out a linkage analysis experiment, numerous genetic markers located at intervals across the whole genome are genotyped in members of the families. These genetic markers can be either STRs or SNPs. Then, a statistical test is performed to examine if there is co-segregation of the disease phenotypes and marker alleles across generations.

To find the disease locus, linkage analysis determines if affected relatives share a DNA segment more often than other segments. The shared segment is more likely to harbor the mutated gene. Genetic linkage refers to markers in a segment having a lower frequency of recombination (i.e. the two markers are rarely separated to

different chromosomes during chromosomal 'crossing-over') and thus being more likely to be inherited together than predicted by chance. The observed level of co-segregation in the pedigree at a particular recombination distance is compared to the likelihood of random segregation, using a likelihood ratio test. The genetic marker locus with the lowest likelihood ratio of recombination is most likely to contain a segment segregating with the disease. Linkage analysis can be applied to pedigrees enriched with participants affected by a monogenic disease. However, for complex traits, linkage analysis has low statistical power to detect polygenic variants.

Another, more comprehensive, tool to identify disease-causing genetic variants is DNA sequencing. If every base in a participant's genome sequence is assessed, then every rare mutation they carry could be identified. Thus, genome sequencing is applicable to detect rare mutations, such as those causing monogenic disorders. The approach has become more popular recently due to technical advances in DNA sequencing and drastically reduced cost. Several mutations causing rare, monogenic forms of myopia have been identified by whole exome sequencing and whole genome sequencing in pedigrees showing monogenic transmission (Guo et al. 2014b; Sun et al. 2015; Jin et al. 2017). In WES, only the 1-2% of the human genome that codes for proteins is sequenced, which makes the technique more cost effective than WGS. However, some disease-causing mutations may be missed by WES if they occur outside of exons.

1.1.5.2. Genome-wide association studies (GWAS)

Genome-wide association studies involve a systematic search through the genome of a large sample of individuals to seek associations between genetic polymorphisms and a trait-of-interest. Typically, GWAS focus on associations between SNPs and phenotypic traits, but the same approach can also be applied to any other genetic

variant type, such as indels, SVs, or STRs.

Human disease GWAS analyses usually adopt the approach known as 'phenotype-first', as opposed to 'genotype-first' methodology, i.e. the participants are recruited by virtue of their clinical manifestations rather than because they are known to have a specific genotype. This method of recruitment is well-suited to a case-control study design, in which individuals with the disease (cases) and without the disease (controls) are analyzed together. Each participant provides a sample of DNA, and millions of genetic variants are read by genetic arrays from these DNA samples. If an allele of a specific genetic variant shows a higher frequency in the case group, the variant is considered to be associated with the disease or trait. The associated variant locations together implicate the gene regions that may influence the risk of disease.

In contrast to methods that focus on a few locations in the genome, GWAS investigate the whole human genome. Compared with genetic linkage studies, GWAS have a higher statistical power to detect weak genetic effects (Risch and Merikangas 1996). GWAS have become the dominant tool for genetic studies of polygenic diseases, to find a set of genetic variants that together confer susceptibility to the disorder (Altmüller et al. 2001).

In a GWA study, a very large number of genetic variants are considered in a GWAS, which causes a multiple testing problem. The very high number of statistical tests increases the likelihood of erroneous inferences, leading to an inflation of the false positive rate, known as a "type 1 error". Hence, a multiple comparison correction is implemented to adjust the level, "alpha" at which a p-value is considered as statistically significant. A Bonferroni correction is a popular method for setting the threshold p-value for statistical significance; this method sets alpha equal to $0.05/n$, where n represents the number of genetic variants evaluated in the GWAS. As a consequence, if testing 1000 genetic variants, then only p-values below 5E-05 would

be considered significant. The family-wise type 1 error rate (FWER) after applying a Bonferroni correction will be 0.05, if all of the tests are independent. For a GWAS, however, many genetic variants are in linkage disequilibrium. This means that some statistical tests will not be independent and lead to a Bonferroni correction being too stringent (FWER < 0.05). As a result, Dudbridge and Gusnanto (2008) have recommended P < 5E-08 as an appropriate p-value threshold for GWAS analyses that test a few million genetic variants. In general, the false discovery rate (FDR) (Benjamini and Hochberg 1995) method is not considered sufficiently stringent to be applied in GWA studies.

1.2. Strabismus

Strabismus is a condition in which the two eyes are not properly aligned and they thus exhibit different directions of gaze. One eye typically fixates straight ahead while the other eye may be directed inwards or outwards, or more rarely, upwards or downwards. The two eyes send different visual images to the brain when one eye is out of alignment. For a young child, the brain learns to ignore the image from the misaligned eye and just attend to the image from the fixating eye. In consequence, the child loses depth perception and the misaligned eye may develop amblyopia (Robaei et al. 2006b).

The commonly used classification for strabismus is based on the direction of the misaligned eye. The most prevalent type of strabismus is convergent strabismus or esotropia, when the deviated eye turns inwards; divergent strabismus or exotropia refers to an outward deviation of the strabismic eye. Also, depending on whether the degree of deviation varies in different directions of gaze, strabismus can be classified into concomitant and incomitant. Patients with concomitant strabismus show the

same degree of ocular misalignment in all directions of gaze; for patients with incomitant strabismus, the degree of misalignment varies depending on the direction of gaze.

1.2.1. Epidemiology of strabismus: prevalence studies

Strabismus is one of the most commonly occurring visual disorders in childhood and adolescence (Lang 1995). Children affected by strabismus may have reduced stereopsis, vision loss and even difficulties in daily life (Satterfield et al. 1993; Sim et al. 2014; Hatt et al. 2016). The prevalence rate of strabismus in children ranges between 2% to 6% (McKean-Cowdin et al. 2013; Hashemi et al. 2015; Griffith et al. 2016; Schuster et al. 2017). Risk factors for strabismus vary depending on the type. Esotropia in childhood is associated with hyperopia and anisometropia; exotropia happens more often in patients with myopia, astigmatism and aniso-astigmatism (Cotter et al. 2011). Other risk factors like prematurity, low birth weight, and smoking during pregnancy are linked to strabismus in childhood, as well (Cotter et al. 2011; Fieß et al. 2017; Schuster et al. 2017).

1.2.2. Epidemiology of strabismus: environmental risk factors

Environmental risk factors also play a role in strabismus development. Comparison between dizygotic twin (DZ) and first-degree relatives revealed different concordance ratios, supporting a role for environmental factors (Wilmer and Backus 2009). (Both DZ twins and first-degree relatives share the same proportion of genetic inheritance, whereas DZ twins typically have a higher degree of shared environment, especially prenatally). Thus, environmental effects potentially explain the greater resemblance of strabismus among DZ twins than first-degree relatives. Although genetic liability is

fundamental to the development of strabismus, other studies have highlighted other environmental risk factors for strabismus, such as low birth weight, prematurity, maternal smoking, and paternal lead exposure (Hakim et al. 1991; Chew et al. 1994; Bremer et al. 1998; Matsuo et al. 2001; Robaei et al. 2006a; Ponsonby et al. 2007). It is also possible that gene-environment interaction contributes to the development of strabismus.

Studies of children with strabismus and their mothers enrolled in the Collaborative Perinatal Project revealed that advanced maternal age, cigarette smoking during pregnancy, and low birth weight (<1500g) contribute to the risk of strabismus (Chew et al. 1994). However, even after accounting for these risk factors, the heritability of concomitant strabismus remained significant (Chew et al. 1995). In a critical review and meta-analysis of strabismus twin studies, Wilmer and Backus (2009) found compelling evidence for a strong genetic influence, but no evidence that environmental factors cause strabismus independently. However, in studies of DZ and MZ twins, Sanfilippo et al. (2012) reported that genetic factors play a major role in eso-deviation strabismus but a lesser role in exo-deviations. Thus, current evidence suggests that the genetic and environmental contribution to strabismus may be specific to the type of deviation.

1.2.3. Genetics of strabismus

Evidence from twin studies support a role for genetics in the aetiology of concomitant strabismus. The concordance rate for strabismus among monozygotic twins is between 73% to 83%, while the rate among dizygotic twins is between 35% to 47% (Paul and Hardage 1994; Matsuo et al. 2002). Family studies have shown that strabismus is associated with a history of parental strabismus (Chen et al. 2020) and have reported a high recurrence risk (‘risk ratio’) among first-degree relatives of a proband with strabismus (Paul and Hardage 1994; Podgor et al. 1996; Parikh et al.

2003).

In the study in a large twin population (1,462 pairs) with clinical data, the heritability of concomitant eso-deviation was estimated to be 64%, yet the heritability for exo-deviation was not significantly different from zero (Sanfilippo et al. 2012). The study by Sanfilippo et al. also examined the genetic correlation between strabismus and refractive error. The addictive genetic correlation for eso-deviation and refractive error was 0.13 and the shared variance was less than 1%, suggesting negligible shared genetic effect (Sanfilippo et al. 2012).

1.2.3.1. Genes causing monogenic strabismus

The first genetic locus associated with concomitant strabismus was identified in a linkage study in 2003 (Parikh et al. 2003). The locus was mapped to chromosome 7p22.1. In 2006, a study examining pedigrees with strabismus syndromes led to the discovery of several candidate gene mutations. These mutated genes played a role in regulating brainstem ocular motoneurons (Engle 2006). Genome-wide screening of non-syndromic strabismus in pedigrees with many affected individuals revealed three susceptibility loci (7p22.1, 4q28.3 and 7q31.2) (Parikh et al. 2003; Shaaban et al. 2009).

1.2.3.2. Genetic variants associated with strabismus

Large scale GWA studies have also found genetic variants associated with strabismus. A GWAS study with 1345 cases with self-reported strabismus and 65,349 controls, identified a locus on chromosome 17 harboring the genes *NPLOC4*, *TSPAN10* and *PDE6G*. Approximately 20 genetic variants in strong linkage disequilibrium were

associated with the phenotype, including 2 candidate causative variants with predictive functional effects (rs6420484 and rs397693108) (Plotnikov et al. 2019). In a separate study, a SNP located within intron 1 of the *WRB* (tryptophan rich basic protein) gene was found to be strongly associated with esotropia (rs2244352) (Shaaban et al. 2018).

1.3. Refractive error

1.3.1. Myopia, hyperopia and astigmatism

Myopia, hyperopia, and astigmatism are three prevalent conditions which affect visual acuity. There has been a rapid increase in the prevalence of myopia in recent decades, especially in East Asia, which has evoked worldwide attention. Researchers are trying to find the causes of myopia and to slow down the speed of its increasing prevalence.

For many years, myopia was recognized as a highly heritable trait. Nevertheless, only recently has significant progress been made in identifying genetic variants associated with refractive error (Tedja et al. 2019). The rapid increase in the prevalence of myopia over the last 30 years implies that the trait is determined not only by genetics effects, but also by environmental or lifestyle factors. Like many other complex traits, common myopia has a complex aetiology that is influenced by an interplay of genetic and environmental factors (Stambolian 2013).

Myopia, or nearsightedness, is a common vision condition in which nearby objects can be seen clearly, but objects farther away are blurry. It occurs when the shape of the eyeball changes, usually due to an increase in the eye's axial length. This causes

light rays entering the eye to be focused in front of the retina, rather than onto the retina. Myopia is commonly defined as a spherical equivalent (SE) refraction of -0.50 diopter (D) or worse. High myopia is usually defined as an SE of less than -6.0 D or an axial length more than 26.0 mm. Myopia - and especially high myopia - is a risk factor for various ocular diseases, including retinal detachment, glaucoma, and cataract (Saw et al. 2005). On the contrary, hyperopia, or farsightedness, is another common vision condition, in which distant objects can be seen clearly, but objects nearby are blurry. In hyperopia, light rays focus behind the retina, typically because the axial length of the eye is too short in comparison to the eye's focal length.

Astigmatism is another type of refractive error in which refract light rays do not focus evenly on the retina, because of a variation of optical power of the eye for light orientated in different directions. Although astigmatism symptoms may be benign, higher degrees of astigmatism in infancy are a risk factor for the development of amblyopia (Brown et al. 2000; Abrahamsson and Sjöstrand 2003). Furthermore, some associations have also been noted between astigmatism and the development of myopia (Fulton et al. 1982; Gwiazda et al. 2000).

Numerous reports have highlighted the link between astigmatism and the development of myopia in children. A connection was found between spherical equivalent and cylindrical refractive error from the association of juvenile-onset myopia with against-the-rule astigmatism (Gwiazda et al. 1993). In a longitudinal study, infants who had against-the-rule astigmatism had an earlier onset of myopia than infants with either with-the-rule or no astigmatism (Gwiazda et al. 1993). Against-the-rule astigmatism in 5- and 6-year-old children was also found to be predictive of later development of myopia by Hirsch (Hirsch 1964), and with faster progression of existing myopia (Grosvenor et al. 1987). In a study of 217 myopic individuals with spherical refractive error of at least -5D or greater in one eye, a moderate correlation was found between the degree of spherical equivalent and

cylinder power ($r = -0.34$, $p < 0.0001$) (Heidary et al. 2005) However, high myopia was identified as a risk factor for the presence of astigmatism, suggesting the direction of causality was unclear (Heidary et al. 2005).

1.3.2. Epidemiology of myopia: prevalence studies

Myopia has become a global public health concern, which affects over 20% of the world's population (Fricke et al. 2018). Myopia increases the risk of serious disorders such as myopic macular degeneration, retinal detachment, glaucoma, and cataract, and is a leading cause of visual impairment and blindness across many countries (Holden et al. 2014). Furthermore, high myopia is emerging as the major cause of blindness in working age individuals of some Asian countries (Iwase et al. 2006; Wu et al. 2011). Recent years have witnessed a surge in cases of myopia worldwide, especially in East Asia (Li et al. 2017), where about 80% of middle school and high school students have myopia (Li et al. 2017). If the current trend maintains its present rate, around 50% of the world will have myopia by the year 2050 (Holden et al. 2016). As announced by the World Health Organization (WHO) in 2012 (Fricke et al. 2012), uncorrected refractive error has a substantial economic impact estimated to be an annual loss of about 202 billion US dollars globally (Fricke et al. 2012). Nowadays, myopia has gained significant attention worldwide, whereas it was once considered as a benign refractive condition.

The highest prevalence rates of myopia are in East and South East Asia, such as in schoolchildren in Singapore, mainland China, Taiwan and South Korea (Xiang et al. 2013; Ding et al. 2017). The myopia prevalence was 65.5% in a group of third year junior high school students (aged 14-15 years, mean 15.25±0.46 years) in Beijing (Li et al. 2017). In South Korea, about 73.0% of children between 12-18 years old are estimated to be affected by myopia (Rim et al. 2016). In Europe, the prevalence rate is lower. One of the largest European studies conducted to date revealed the

prevalence rate was 42.7% in a French cohort aged 10 to 19 (Matamoros et al. 2015). In the United Kingdom, the prevalence rate is approximately 2% in 6-7-year-olds and 15% in 12-13-year-olds. The lowest prevalence rate has been found in Africa, where in one study published in 2003, only 4% of 5-to-15-year-olds were myopic (Naidoo et al. 2003).

Multi-ethnicity studies have found a different prevalence of myopia among various ethnic groups. Individuals of Chinese ethnicity typically have the highest prevalence rate for myopia. For instance, in Singapore, individuals of Chinese ethnicity had a higher odds ratio for myopia and high myopia versus individuals of Malay and Indian ethnicity (Pan et al. 2013). A study from Australia found 39.5% of East Asian children had myopia, which was much higher than that of European children (4.6%) and Middle Eastern children (6.1%) (Ip et al. 2008). Also, in United States, the highest myopia prevalence rate (18.5%) was found among Asians, followed by Hispanics (13.2%) and Africans (6.6%), while white Americans (4.4%) had the lowest myopia rate (Kleinstein et al. 2003).

The primary reason cited for the increase in myopia prevalence in East and Southeast Asia over recent decades is increased educational pressure (Baird et al. 2020; Morgan et al. 2021). Genetic variants associated with refractive error are largely shared across European and East Asian individuals (Tedja et al. 2018a), however schooling is often more intensive and extra-curricular evening classes are more common in countries such as China, Singapore and Hong Kong compared to the West. Children in East and Southeast Asia also spend limited time outdoors, compared to children in Europe and the United States, with reduced exposure to high light intensities from sunlight and far-distance vision (Morgan et al. 2021).

1.3.3. Epidemiology of myopia: environmental risk factors

The level of education attainment is an important environmental risk factor for myopia. An association between myopia and education attainment was first found in the early 1900's (Harman 1913). Later, numerous studies investigated the association of educational and myopia development. Strong associations between educational exposure and myopia have been found in epidemiological studies, showing a higher prevalence rate of myopia among individuals holding a university degree compared to those only completing a secondary or primary school education(Au Eong et al. 1993; Cumberland et al. 2015). Through the *Mendelian randomization* (MR) approach, studies suggested there is a causal relationship between years of schooling and myopia (Cuellar-Partida et al. 2016; Mountjoy et al. 2018).

Near work is another environment factor investigated in myopia studies. The term *near work* refers to the activities performed at a short distance, such as reading, writing, watching TV, and playing video games (Mutti et al. 2002). More near work exposure was found to be associated with a higher prevalence of myopia in children (Guo et al. 2016). Another study in adults revealed that prolonged near work was associated with myopia development and axial elongation (Woodman et al. 2011).

Time spent outdoors or outdoor activity has been regarded as an important factor for myopia control. In 12-year-old children, students with higher levels of outdoor activities (sport and leisure activities) had more hyperopic refractions and a lower myopia prevalence (Rose et al. 2008). A study in China recruited 1,903 primary school students for a cluster randomized 'intervention-control' study. Students in the intervention group were given an addition 40 minutes outdoor activity per school day. This intervention caused a significant reduction in the cumulative myopia incidence rate between the intervention group and the control group (30.4% vs. 39.5%, respectively, $p<0.001$) over the next three years (He et al. 2015). There are several theories concerning the mechanism underlying the protective effects of outdoor activities on myopia development, but to date the true mechanism remains

unknown.

The association of refractive errors with working distances was also investigated. Hartwig et al. (2011) hypothesized that head and eye movements were different between myopes and non-myopes while doing near work, and measured the horizontal and vertical amplitude of eye movements and head movements while subjects performed three near tasks. No significant difference in the eye movement or head movement was found in the same task between myopes and emmetropes (Hartwig et al. 2011).

Few studies have addressed the association between myopia and near-working distance compared to its association with the duration of near work (Saw et al. 2002; Lu et al. 2009; Huang et al. 2015; Philipp et al. 2022). Those studies that have been performed have mainly relied on self-reported working distance (Wu et al. 2015), which may have reduced the statistical power to detect any association.

1.3.4. Genetics of myopia

1.3.4.1. Genes causing monogenic high myopia

To date, 26 myopia loci have been discovered via linkage analyses, and numerous candidate genes inside the linkage regions have been analyzed. Of these 26 loci, 23 are found on autosomal chromosomes, and the other 3 on the X-chromosome (Table 1.2) (Cai et al. 2019; Ouyang et al. 2019).

Locus	OMIM	Location	Inheritance	Related gene	Myopia severity	References	Causality
MYP1	310460	Xq28	X-linked	OPN1LW	−6 to −23 D (mean, −8.48 D)	(Guo et al. 2010); (Ratnamala et al. 2011); (Orosz et al. 2017)	Not confirmed.
MYP2	160700	18p11.31	AD	TGIF	−6 to −21 D	(Young et al. 1998b)	Controversial.
MYP3	603221	12q21-q23	AD	DCN, LUM, DSPG3	−6.25 to −15 D	(Young et al. 1998a); (Wang et al. 2017); (Okui et al. 2016); (Park et al. 2013)	Controversial.
MYP6	608908	22q13	AD	SCO2	no less than −1.00 D	(Stambolian et al. 2004); (Tran-Viet et al. 2013)	Controversial.
MYP7	609256	11p13	Multifactorial	PAX6	−12.12 to +7.25 D	(Hammond et al. 2004)	Not confirmed.

MYP17, MYP4	608367	7p15, 7q36	AD	VIPR2	> -6 D	(Naiglin et al. 2002); (Paget et al. 2008); (Klein et al. 2007); (Ciner et al. 2008)	Not confirmed.
MYP21	614167	1p22.2	AD	ZNF644	−6.27 to −20 D; > -6 D	(Shi et al. 2011a);(Tran-Viet et al. 2012)	Replicated.
MYP23	615431	4p16.3	AR	LRPAP1	> -17 D	(Aldahmesh et al. 2013)	Replicated.
MYP5	608474	17q21-q22	AD	COL1A1, CHAD	−5.5 to −50 D (mean, −13.93 D)	(Paluru et al. 2003a)	Not confirmed.
MYP8	609257	3q26	Multifactorial	NR	−12.12 to +7.26 D	(Hammond et al. 2004)	Not confirmed.
MYP9	609258	4q12	Multifactorial	NR	−12.12 to +7.27 D	(Hammond et al. 2004)	Not confirmed.
MYP10	609259	8p23	Multifactorial	NR	−12.12 to +7.28 D	(Hammond et al. 2004)	Not confirmed.

MYP12	609995	2q37.1	AD	SAG, DGKD	−7.25 to −27 D	(Paluru et al. 2005)	Not confirmed.
MYP14	610320	1q36	NR	NR	−3.46 D (mean)	(Hammond et al. 2004)	Not confirmed.
MYP15	612717	10q21.1	AD	CDH15, ZWINT	-7 D (mean)	(Nallasamy et al. 2007)	Not confirmed.
MYP11	609994	4q22-q27	AD	RRH	−5 to −20 D	(Zhang et al. 2005)	Not confirmed.
MYP13	300613	Xq23-q27.2	X-linked	NR	−6 to −20 D; < -7 D	(Zhang et al. 2006); (Zhang et al. 2007)	Not confirmed.
MYP16	612554	5p15.33-p15.2	AD	NR	> -6 D	(Lam et al. 2008a); (Lam et al. 2008b)	Not confirmed.
MYP18	255500	14q22.1-q24.2	AR	NR	> -6 D	(Yang et al. 2009)	Not confirmed.
MYP19	613969	5p15.1-p13.3	AD	CDH6, CDH10, CDH12, PDZD2, GOLPH3	< -6 D	(Ma et al. 2010)	Not confirmed.
MYP20	614166	13q12.12	AD	MIPEP, C1QTNF9B-AS1,	< -6 D	(Shi et al. 2011b)	Not confirmed.

				C1QTNF9B			
MYP22	615420	4q35.1	AD	CCDC111	< -6 D	(Zhao et al. 2013)	Controversial.
MYP24	615946	12q13.3	AD	SLC39A5	< -6 D	(Guo et al. 2014a)	Replicated.
MYP25	617238	5q31.1	AD	P4HA2	−6 to −20 D	(Guo et al. 2015)	Replicated.
MYP26	301010	Xq13.1	X-linked	ARR3	> -6 D	(Xiao et al. 2016)	Not confirmed.
MYP27	606027	8q24.3	AD	CPSF1	< -6 D	(Ouyang et al. 2019)	Not confirmed.

The first genetic locus linked with myopia was reported in 1990 from the study of a family affected with Bornholm Eye Disease, and the genetic locus was named "MYP1" (Schwartz et al. 1990). Other studies of pedigrees affected by the syndrome have confirmed linkage to the MYP1 locus (Guo et al. 2010). Ratnamala et al. (2011) demonstrated X-linked recessive inheritance at the MYP1 locus in a pedigree with non-syndromic high myopia. Recently, mutations in *OPA1LW* were reported to be responsible for Bornholm Eye Disease, as well as the form of non-syndromic high myopia that mapped to MYP1 (Orosz et al. 2017). In 1998, Young et al. (1998b) identified the MYP2 locus with an autosomal dominant pattern of high myopia via a pedigree study. MYP3 was found in a large family of Greek/Italian ancestry with autosomal dominant pattern (Young et al. 1998a). Decorin (*DCN*), Lumican (*LUM*) and Dermatan sulfate proteoglycan-3 (*DSPG3*) were identified as candidate genes; these genes code for proteoglycans that play a role in the extracellular matrix organization of the sclera. However, the *LUM* gene was demonstrated not to be associated with myopia in genetic analyses of the Korean, Japanese and Chinese population samples (Park et al. 2013; Okui et al. 2016; Wang et al. 2017).

A study of American families of Ashkenazi Jewish descent identified the MYP6 locus on 22q12 (Stambolian et al. 2004). In 2013, Tran-Viet et al. (2013) found the *SCO2* gene to be the disease-causing gene at the MYP6 locus . A twin study in 506 DZ and MZ twin pairs was conducted to assess the heritability of refractive error, and a genome-wide linkage scan was performed in 221 DZ twin pairs (Hammond et al. 2004). Four novel loci were reported in this study, MYP7-MYP10. A regression-based quantitative trait loci (QTL) linkage study in Ashkenazi Jewish families identified a novel locus, designated MYP14, on chromosome 1p36 (Wojciechowski et al. 2006). A genome-wide scan for high myopia in 23 families identified the MYP4 locus on chromosome 7q36 (Naiglin et al. 2002). Paget et al. (2008) found no linkage to chromosome 7p36 in a follow-up linkage analysis. However, a non-parametric model

suggested significant linkage to nearby 7p15. Later studies also found the loci re-mapped to 7p15 (Klein et al. 2007; Ciner et al. 2008), so the MYP4 name was replaced by MYP17.

Paluru et al. (2003a) identified the MYP5 locus on chromosome 17q21-q22 in a multigenerational large English/Canadian family with familial high myopia. Chondroadherin (*CHAD*) and the extracellular matrix protein *COL1A1* were proposed as candidate genes: the latter gene regulates the process of collagen fibril assembly and affects numerous tissues, including the sclera. The MYP12 locus on chromosome 2q37.1 was identified in an American family of European ancestry. Candidate genes were sequenced but no causative mutations were found (Paluru et al. 2005). Linkage analysis of high myopia in a large Hutterite family from South Dakota, identified the MYP15 on chromosome 10q21.1 (Nallasamy et al. 2007).

The majority of the loci associated with monogenic forms of myopia listed in Table 1.2 have an autosomal dominant inheritance pattern, except for MYP18, MYP23, and MYP26. Yang et al. identified an autosomal recessive locus for high myopia on chromosome 14q22.1-q24.1 (MYP18). The causal gene was not found (Yang et al. 2009). The MYP23 locus, which was mapped to chromosome 4p16.3, was associated with non-syndromic high myopia. Exome sequencing revealed *LRPAP1* as the causative gene (Aldahmesh et al. 2013). Xiao et al. (2016) reported the MYP26 locus on chromosome Xq13.1 after studying 3 Chinese families affected with female-limited early-onset high myopia. A heterozygous mutation in *ARR3* was identified as being responsible for this form of X-linked, female-limited high myopia.

The most recent discovery of a locus for non-syndromic high myopia was in 2019, when a mutation in the *CPSF1* gene was reported to cause early-onset high myopia (MYP27) (Ouyang et al. 2019).

1.3.4.2. Genetic variants associated with refractive error and myopia

GWA studies have been conducted for myopia to identify genetic variants associated with the phenotype. A common design of GWAS for myopia is the case-control study design, i.e. a binary trait classification of cases with myopia vs. non-myopic controls. A GWAS using a two-stage case-control design was conducted for pathological myopia in Japanese participants (Nakanishi et al. 2009). In the first stage, 411,777 SNPs were evaluated in a GWAS with 297 high myopia cases and 934 controls. In the second stage, 22 SNPs with a P-value smaller than 0.0001 in the first stage were tested for association in a replication sample of 533 cases and 977 controls. Finally, SNP rs577948 was found to be associated with high myopia (p = 2.22E-7, OR =1.37, 95% CI 1.21-1.54). The SNP is located at chromosome 11q24.1, nearby two candidate genes, *BLID* and *LOC399959*. The two genes are situated within 200-kb of rs577948, and both were demonstrated to be expressed in human retinal tissue.

In a study of 520 cases with high myopia and 520 controls in Japan, 39 SNPs located on 21q22.3, which were previously reported to be associated with high myopia, were tested using a chi-squared test and Fisher's exact test (Nishizaki et al. 2009). One SNP (rs2839471) located within the *UMOLD1* gene was significantly associated with the disease.

In a meta-analysis of two GWA datasets in Singaporean Chinese participants and a follow-up replication cohort from Japan, two SNPs (rs12716080 and rs6885224) in the gene *CTNND2* on chromosome 5p15 were found to be associated with high myopia (Li et al. 2011). Variant rs6885224 was found to be significantly associated with high myopia in an independent replication sample.

Another study design is to use refractive error as a continuous trait and to search for associated genetic variants. Two separate GWA studies conducted in Europe (Hysi et al. 2010; Solouki et al. 2010) found two associated loci for myopia on chromosome 15q14 and 15q25, near the *GJD2* gene and the *RASGFR1* gene. In 2010, the Consortium for Refractive Error and Myopia (CREAM) was established. In 2013, the CREAM consortium conducted a genome-wide meta-analysis of 37,382 individuals from 27 studies of European ancestry, and 8,376 individuals from 5 Asian cohorts. A total of 16 novel associated loci for refractive error were identified, in addition to the *GJD2* gene and *RASGFR1* gene variants (Verhoeven et al. 2013b).

A GWA study performed by the personal genomics company 23andMe Inc., analyzed data for 45,771 individuals of European ancestry (Kiefer et al. 2013). A survival analysis for age-of-onset of myopia identified 22 significant associations ($p < 5E-08$), including 20 novel discoveries. Ten of the 20 novel associations were replicated in a separate cohort of 8,323 participants that reported whether their age-of-onset of myopia was before 10 years or not. These 22 associations in total explained 2.9% of the variance in myopia age-of-onset. It should be noted that the ethics of direct-to-consumer genetic testing, as offered by 23andMe, has been questioned (Hsu et al. 2009).

Recently, in a meta-analysis of GWAS involving 542,934 individuals of European ancestry, 904 independent SNPs were found significantly associated with refractive error. This work identified 336 novel genetic loci (Hysi et al. 2020).

Other ocular phenotypes related to myopia have also been studied using the GWAS approach, including GWAS for corneal and refractive astigmatism (Shah et al. 2018), ocular axial length (Fan et al. 2012; Cheng et al. 2013), and macular thickness

(Hosoda et al. 2018; Gao et al. 2019).

1.3.5. Pathological complications of myopia

Myopia is associated with complications that can lead to irreversible visual impairment later in life. Complications including myopic macular degeneration (MMD), retinal detachment (RD), cataract, and glaucoma have been found to be association with myopia (Ohno-Matsui et al. 2016).

For myopes, especially people suffering from high myopia, MMD is the most common detrimental complication. MMD includes pathological features such as lacquer cracks, Fuchs spot, choroidal neovascularization (CNV), and chorioretinal atrophy (Ohno-Matsui et al. 2015). RD is one of the most sight-threatening peripheral retinal lesions in high myopes (Lam et al. 2005; Verhoeven et al. 2015), alongside other features including pigmentary degeneration and paving stone degeneration (Lam et al. 2005; Verhoeven et al. 2015).

High myopia (HM) (defined as a spherical equivalent less than -6.0 D or an axial length more than 26.0 mm). An earlier age-of-onset of myopia (Jensen 1995; Chua et al. 2016; Hu et al. 2020) and a higher baseline spherical equivalent (SE) refractive error in childhood (Gwiazda et al. 2007) are highlighted to be predictive of HM later in life. The Singapore Cohort Of the Risk factors for Myopia (SCORM) study has found, among 11 years old children, the age-of-onset of myopia and the duration of myopia are the most important predictive factors for HM in later childhood (Chua et al. 2016). In the SCORM study, astigmatism was not identified as a risk factor for the severity of myopia later in life among myopic children. Another study in Denmark showed similar results, where those with a younger age-of-onset of myopia (9-12

years as baseline) were more likely to develop HM after 8 years of follow-up (17-20 years old) (Jensen 1995).

Chapter 2. Methods

2.1. UK Biobank

UK Biobank recruited over 500,000 participants aged 40-69 years during 2006-2010, with the aim of performing a detailed investigation of the genetic and nongenetic factors contributing to the diseases of middle-aged and elderly people (Sudlow et al., 2015). Prospective participants were invited to attend an assessment, where they completed an automated questionnaire and were interviewed about lifestyle, dietary habits, and medical history; basic traits such as weight, height, and blood pressure were measured; and blood, urine, and saliva samples were collected. DNA and other metabolites were extracted from these samples. Genome-wide array-based genotyping and whole exome sequencing were carried out. An eye and vision component was introduced into the UK Biobank baseline assessment visit beginning in 2009; this was towards the end of the recruitment period of 2006-2010. In total, 117,649 people took part in the eye and vision assessment. The eye and vision tests comprised of a modified logMAR visual acuity test on a computerized system, autorefraction and keratometry (Tomey RC-5000), as well as measurements of intraocular pressure, corneal hysteresis and corneal resistance factor (Reichert ORA Ocular Response Analyser). Colour retinal fundus photography, together with spectral domain optical coherence tomography (SD-OCT) in both eyes, were undertaken within a smaller group (Topcon 3D OCT-1000 Mark II) (68,151).

The UK Biobank obtained ethics approval from the North West Multi-centre Research Ethics Committee (Reference No. 06/MRE08/65) and complied with the principles of the Declaration of Helsinki. All participants provided informed written (digital) consents.

UK Biobank is one of the largest and most inclusive population cohort studies globally. As an easily accessible database of deep phenotypic, genomic, imaging, and

health outcomes, UK Biobank provides excellent opportunities for eye and vision studies (Chua et al., 2019). The large sample size and standardized measurements permit researchers to detect and quantify small effects. However, some features of the UK Biobank may serve as limitations. For instance, the fact that many disease statuses were self-reported by participants could introduce inaccurate classification error (Shweikh et al., 2015). Meanwhile, the participants in the UK Biobank study were more likely to live in socioeconomically advantaged regions, where people have better educational level, are less likely to experience tobacco or alcohol addictions, and have lower obesity rates (Fry et al., 2017). The studies are, therefore, not fully representative of the general UK population. Furthermore, the response rate of the UK Biobank was only 5.5%, which confirms its limited external validity. With respect to certain diseases, this could have resulted in under-ascertainment (Shweikh et al., 2015).

2.1.1. Phenotypes in UK Biobank: Eye and Vision-related Data

Moorfields Eye Hospital and the UCL Institute of Ophthalmology in London developed the eye and vision module. The data collection procedure for the eye examinations received core funding from the Wellcome Trust, The Medical Research Council and The Department of Health. Valid eye data from 117,649 participants were available within UK Biobank. The longitudinal design of the study and the accessible link to other UK Biobank data make it one of the most valuable resources for ophthalmology research in the UK.

Non-cycloplegic refractive error was carried out by using the Tomey RC 5000 autorefractor (Tomey Corp., Nagoya, Japan). The autorefractor recorded up to ten refractive error measurements for each eye, along with a reliability score, ranging

from 0 to 9 (smaller scores mean more reliable measurements). Scores of $\leqslant$4 were considered reliable. The average of all available reliable scores was calculated as the final refractive error of that eye.

Spherical equivalent refractive error was calculated as sphere power plus half the cylinder power. The refractive error of an individual was taken as the average spherical equivalent of the two eyes (Tedja et al. 2018b). The threshold for myopia was set as a spherical equivalent refractive error $\leqslant$ − 0.50 diopters (D). Refractive astigmatism was taken as the average cylinder power between the two eyes. The status of individuals with and without astigmatism ≥ 1.00 D was recorded as a binary variable (Shah et al. 2018). Anisometropia was calculated as the difference in spherical equivalent between the two eyes. The status of individuals with and without anisometropia $\geqslant$ 1.00 D was recorded as a binary variable (Qin et al. 2005).

The type of strabismus, for instance esotropia and exotropia, was not ascertained through the data assessment procedure, so the 'self-reported' strabismus information for UK Biobank participants consisted of all plausible subtypes of strabismus. Similarly, information about strabismus surgery during childhood was not available due to the lack of collection for medical records on hospital in-patient operations until April 1997. In addition, patients who had strabismus, but did not require any prescription glasses or contact lenses because of the mildness of condition, would have been classified into 'control' group instead of 'case' group for the self-reported strabismus.

A positive angle Kappa can lead to a misdiagnosis of convergent strabismus in clinical studies. However, in the UK Biobank study, strabismus status was based on participant self-report. Therefore, angle Kappa is a less likely to bias the diagnosis of strabismus in the current work.

2.1.2. Genetic Data in UK Biobank

Whole-exome sequencing data facilitate the direct assessment of protein-altering variants. Such variants are more likely to have functional effects and therefore can be more readily interpreted than non-coding variants. Thus, analysis of WES data increases the likelihood of identifying pathways, disease mechanism, and has the potential to be used in therapeutic target discovery and validation (Cohen et al. 2006; Scott et al. 2016; Dewey et al. 2017; Abul-Husn et al. 2018) and in precision medicine (Abul-Husn et al. 2016).

The UK Biobank released WES data for 200,000 participants, which included 1,135 parent-offspring pairs, 3,855 full-sibling pairs, including 101 trios, 27 monozygotic twin pair and 7,461 second degree genetically determined relationships. In March 2019, the first release of WES data for 50,000 UK Biobank participants was made available. These participants were prioritized for the acquirement of whole body MRI imaging data, enhanced baseline measurements, hospital episode statistics (HES), and/or linked primary care records. Data for an additional 150,000 participants was made available in October 2020 (UK Biobank 2022a).

To generate the WES data, DNA from the exome was captured by an IDT xGen Exome Research Panel v1.0 including supplemental probes. The basic design targeted 39 Mbp of the human genome (19,396 genes). The variant callset included variants in both the target regions and 100 bp flanking regions upstream and downstream of each capture target (UK Biobank 2022b). The location of targeted regions is coordinated to Genome Reference Consortium Human Build 38 (GRCh38). Primary and secondary analysis for the UKB 200k release were performed with an updated

Functional Equivalence (FE) protocol that retained original quality scores in the CRAM files (referred to as the "OQFE protocol") (UK Biobank 2022a). The OQFE protocol aligns and duplicate-marks all raw sequencing data (FASTQs) to the full GRCH38 reference, and further generates a single multi-sample VCF (pVCF) file for all 200,000 samples. PLINK files were directly generated from this pVCF file. No variant- or sample-level filters were pre-applied to the pVCF or PLINK files, but the pVCF file contains allele-read depths and genotype qualities for all genotypes, onto which quality control metrics and analysis-specific filters can be applied.

2.2. Statistical analyses

This section describes the statistical methods and statistical software used in this thesis.

2.2.1. Linear Regression

Regression modeling is one of the most important statistical approaches applied into analytical epidemiology (Bender 2009), as well as for genome-wide association studies (Bush and Moore 2012). The most frequently asked question in analytical epidemiology is whether an exposure such as hours of reading has an impact on a response, such as myopia. Similarly, the most frequently asked question in a genome-wide association study is whether a genetic variant such as a single nucleotide polymorphism has a significant association with a phenotypic trait such as refractive error. Exposure variables are usually measured as categorized or continuous indicators, while the genetic variants are usually coded based on the different alleles.

A common aim of cohort studies and case-control study designs in epidemiology is to investigate the relations between an exposure and an outcome (Bender 2009).

Regression analysis can – potentially – control for confounding effects. The association between an exposure (such as a genotype) and an outcome response can be biased by confounding effects in non-randomized studies. (Generally, in epidemiology, it is impossible to randomly assign participants to different exposure groups). To statistically control for the effects of confounding factors, multivariate regression is used to model the relationship between a dependent variable and one or multiple explanatory variables, also called independent variables, predictor variables, covariates, or risk factors.

For example, a linear regression can be applied to study the effects of time spent on near work activities on the refractive error averaged between the two eyes. The fundamental regression equation is given by:

$$Y = \beta_0 + \beta_1 x + e$$

Where β_0 is called the intercept, β_1 is the regression coefficient for the dependent variable x, and e is called the residual (which describes the deviation of each individual i from the mean of Y given $x = x_i$). A straight line is fitted through the model by ordinary least squares, where the intercept β_0 represents the mean value of Y when $x = 0$ and regression coefficient β_1 represents the slope of the straight line, denoting the average increase of Y for a one-unit increase of the independent variable x. Two fundamental assumptions should be fulfilled to yield interpretable results: the first assumption is that the expected value of Y is a straight-line function of x. The second assumption is that the residual term is normally distributed with mean 0 and variance σ^2, which is the same for any value of x.

In a GWAS, the regression coefficient β_1 is the average change in the trait of interest per copy of the 'risk' allele (the allele that confers a risk of developing the disease).

2.2.2. Logistic Regression

Logistic regression is used when the dependent variable is a binary variable (e.g., self-reported strabismus, yes/no). Logistic regression can be applied in both cohort studies and case-control studies (Anderson 1972; Mantel 1973; Huang et al. 2015), which makes logistic regression one of the most important statistical models in epidemiologic and genome-wide association studies.

The basic question addressed by logistic regression is, "Is there an association between the independent variable x and the dependent binary variable Y. To overcome the inconvenient restriction of the binary value (0 or 1) and create a continuous variable that spans the whole number domain, a mathematically convenient term $\log\left(\frac{\pi}{1-\pi}\right)$, called the logit, is derived, in which π corresponds to the event probability $\pi = P(Y = 1)$ and $\left(\frac{\pi}{1-\pi}\right)$ is the "odds" of the event occurring. Thus, the logit is the natural logarithm of the odds. It can take a value from negative to positive infinity, as the dependent variable ranges from zero to one. The logistic regression is given by the linear relationship between the logit and the value of independent variables,

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

The $\log\left(\frac{\pi}{1-\pi}\right)$ represents the natural logarithm of odds, π is the probability of the binary outcome being 1, β_0 is the intercept, and β_1 is the natural logarithm of the odds ratio of a one unit increase in x, and e is the residual term.

The effect size of X_j on the dependent variable in the logistic regression is best described by means of $\exp(\beta_j)$, because in the logistic regression model,

$$\exp(\beta_j) = OR_j$$

Where OR_j represents the odds ratio for X_j adjusted for the other independent variables. In the case of a continuous explanatory variable X_j (and a model without interactions), the OR_j describes the factor by which the odds of an event changes for each one-unit increase in X_j.

In a highly unbalanced case-control association study (i.e. where the number of controls is much greater than the number cases) or in studies of rare genetic variants, the standard logistic regression estimate can be biased (Ma et al. 2013). Instead, a bias-corrected estimate can be performed by a Firth logistic regression (Firth 1993; Wang 2014). Firth logistic regression relies on a more effective score function to counteract the asymptotic expansion of the bias of the maximum likelihood estimation. For generalized linear models such as in logistic regression, Firth's approach is equivalent to utilizing Jeffreys invariant prior to penalize the likelihood function (Firth 1993). Firth's approach achieves bias-reduction for small-scale samples, as well as providing finite and reliable estimates even in analysis of association with rare variants (Wang 2014).

2.2.3. Covariates

The age of participants when they underwent their baseline tests at a UK Biobank assessment centre, refractive error averaged between the two eyes (avMSE), and the first 10 ancestry principal components (PC) were included as quantitative covariates. Sex was coded as binary variables. PC analysis is a dimension reduction technique. When applied to high-density genotype data, the major PCs (PCs 1-10) have been found to correspond to quantitative estimates of genetic ancestry. For example, the first 2 PCs can be used to distinguish individuals of African, Asian and European ancestry. Accordingly, the major PCs can be used to identify samples with similar ancestry or to control for population structure in association studies.

2.2.4. Chi-squared and Fisher's Exact Test

The Chi-squared test can be applied to test the independence of two categorical variables. By comparing two variables in a contingency table, the Chi-squared test is applied to determine whether the distribution of the variables differ from each other. In the standard application of the Chi-squared test, the observations are classified into mutually exclusive classes. The null hypothesis is that there are no differences between the classes in the population. A Chi-squared statistic computed from the observations follows a Chi-squared distribution. The Chi-squared test evaluates the likelihood that the null hypothesis is true based on the observations.

The formula for the Chi-squared statistic used in the Chi-squared test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ_c^2 is the Chi-squared statistic, and c is the degree of freedom. O_i is the observed value and E_i is the expected value. The summation symbol is for the data in every single cell i of the contingency table. The degrees of freedom for the contingency table has the formula shown as:

$$degree\ of\ freedom = (number\ of\ rows - 1)(number\ of\ columns - 1)$$

According to the Chi-squared test value and degrees of freedom, the *p*-value for the Chi-squared test could be found in the Chi-squared table. In general, a *p*-value lower than 0.05 can be used to reject the null hypothesis.

In the epidemiology and genetic studies, a Chi-squared test can be applied to investigate whether or not a genetic variant or exposure of interest is associated with the phenotype. For a case-control study, the Chi-squared test contingency table would classify data by case/control condition as one variable and the allele or exposure condition as the other variable. An example contingency table is shown in Table 2.1:

Table 2.1 Example contingency table for case-control study

	case	control	sum
Exposure	a	b	m_1
No exposure	c	d	m_2
sum	n_1	n_2	S

a, b, c, d are the numbers of observations for each single cell, and n_1, n_2, m_1, m_2 are the sum of each column and row, and S is the total number of samples. The purpose of the test is to determine whether the null hypothesis of no association between the exposure and phenotype is true. For this 2x2 contingency table, the degree of freedom is 1, the expected number under the null hypothesis can be calculated from the table: for example, the expected number in the cell for case-exposure sample should be $\frac{n1m1}{S}$. By using the observed and expected data sets, the Chi-squared statistic with 1 for degree of freedom can be calculated. The corresponding p value can be obtained by looking up the Chi-squared statistic in a probability table. A higher value of the Chi-squared statistic means a lower correlation between the observed data and the expected data set and a lower p value, which indicates the violation of the null hypothesis that no association exists between the exposure and the phenotype.

In genetics, the Chi-squared test can be applied to study the distribution of alleles. Alleleic disparities in a population could arise by chance, or from external factors. The external factors are mainly from environmental effects, which contribute to the statistically significant difference between the observed data set and the expected data set. If the p value calculated from Chi-squared test is lower than the threshold set by the researcher, then the null hypothesis of no association would be rejected and it can be determined that the variance is due to more than chance, i.e. an external factor has contributed to that variance.

In the genome-wide association study, the Chi-squared test can be applied to pinpoint the locus associated with a categorical phenotypic (e.g., case/control labels for individuals). To comply with the requirement of the Chi-squared test, both genotypes and phenotypes have to be categorical variables. Due to the ease of

genotyping and abundance across genomes, biallelic SNP genotypes are usually used as the genetic variants for association studies. For biallelic SNPs, minor allele homozygous, heterozygous, and major allele heterozygous sites are coded as 0, 1, and 2. A contingency table for the Chi-squared test can be created by the different allele/phenotype combinations.

In practice, Fisher's exact test can be applied as an alternative to the Chi-squared test for testing the associations in small or unbalanced datasets. When the sample sizes are small, or the data are very unequally distributed among the cells of the contingency table, the sampling distribution of the test statistic is an inadequate approximation of the Chi-squared distribution. Such scenarios may lead to biased conclusions concerning the hypothesis of interest (Mehta et al. 1984). In contrast, as long as the experimental procedures keep the row and column totals fixed, Fisher's exact test can be used regardless of the sample characteristics.

In a GWAS, statistical hypothesis testing is based on rejecting the null hypothesis of no association if the observed *p*-value is 'low'. Testing multiple hypotheses increases the likelihood of incorrectly rejecting the null hypothesis for some of the tests, i.e. making type 1 errors or reporting 'false-positive' findings. A Bonferroni correction is a commonly-applied method to control the type 1 error rate of the full analysis such that 'experiment-wide' type 1 error rate is below a specified threshold (usually $p < 0.05$).

In the weighted Bonferroni method, different weights are assigned to two or more different endpoints, with the sum of the weights equal to 1.0 (e.g., 0.4,0.4,0.2 for three endpoints). These weights are pre-set in the design of trials, taking into consideration the importance of the different endpoints, the likelihood of success, or other factors. One of the most popular ways to perform the weighted Bonferroni test

is by assigning a specific 'amount' of alpha (e.g. 0.05) to each endpoint. The weighted Bonferroni method is often applied by dividing the overall alpha into unequal portions. By multiplying the overall alpha by the assigned weighting factors, the sum of the endpoint-specific alphas will remain as the overall alpha, such that each calculated *p*-value is compared to the assigned endpoint-specific alpha. By introducing prior information to assign specific weights, the weighted Bonferroni correction not only controls the family-wise error rate, but can have higher power than the standard Bonferroni procedure (Rubin et al. 2006; Wasserman and Roeder 2006).

Rubin *et al*. (2006) and Wasserman and Roeder (2006) proposed a weighted Bonferroni procedure that used optimal weighting factors. Under the assumption that the mean of all test statistics are known, these optimal weights were calculated by maximizing the average power of the weighted Bonferroni correction.

Chapter 3. Chromosome 17 Association Study for Strabismus

3.1. Introduction

Strabismus is a common condition characterized by constant or intermittent abnormal alignment of the eyes that leads to loss of binocular vision (Chapter 1.2.).

Strabismus is known to be affected by both the genetic and environmental effects. Several studies have been performed to identify the genetic risk factors associated with strabismus (Graeber et al. 2013; Kruger et al. 2013; Maconachie et al. 2013). There is evidence that the underlying causal variants have complex effects on the risk of developing strabismus (Sanfilippo et al. 2012; Ye et al. 2014). Meanwhile, other lines of evidence support the hypothesis that there are rare, monogenic forms of strabismus. For example, strabismus was found to co-segregate with genetic markers at 7p22.1 in a linkage analysis study of families with a pathological history of Mietens-Weber Syndrome and Lamb-Shaffer Syndrome (Parikh et al. 2003). Recent GWA studies identified a small number of common genetic variants that confer susceptibility to strabismus. Shaaban et al. (2018) performed a GWAS and identified a single variant (rs2244352; OR = 1.33, p = 9.58E-11) that was significantly associated with non-syndromic strabismus. Plotnikov et al. (2019) reported approximately 20 variants in almost perfect linkage disequilibrium (LD) across the *NPLOC4–TSPAN10–PDE6G* gene cluster (lead variant: rs75078292; OR = 1.26, p = 2.24E−08) strongly associated with self-reported strabismus.

Here, the hypothesis was tested that a commonly occurring STR polymorphism located on chromosome 17 confers susceptibility to non-syndromic strabismus. An STR-based chromosome-wide association study for self-reported strabismus was carried out. The study focused attention on chromosome 17 because a previous GWAS in UK Biobank participants by Plotnikov et al. (2019) had identified a large-

effect sized association on chromosome 17. In view of the strong association between strabismus and refractive error, the average mean spherical equivalent (avMSE) was included as a covariate in the analysis in order to avoid the cofounding effects of genetic variants associated with refractive error.

3.2. Methods

3.2.1. Selection of Participants

The genetic analysis and subsequent analyses were restricted to unrelated UK Biobank participants of European ancestry who were part of the October 2020 (WES 200k) data release. The genetic ancestry principal components (PCs) provided by Bycroft et al. (2018) were used to define a cluster of individuals with European ancestry. Participants whose genetic ancestry PCs did not cluster with Europeans were excluded. Only individuals who did not withdraw their consent were studied. Moreover, participants with a mismatch between their self-reported and genetically-inferred sex, or whose imputed genotype data showed high heterogeneity (heterozygosity >4 standard deviations from the mean level), were also excluded. After applying these filters, there were n=181,170 individuals remaining. Next, participants who self-reported a history of eye trauma resulting in loss of vision, cataract surgery, laser eye surgery or corneal graft surgery were excluded, as were individuals whose hospital records (ICD10 codes) indicated a history of cataract surgery, eye surgery, retinal surgery, or retinal detachment surgery. Participants without a valid mean spherical equivalent autorefraction measurement or who were recruited from a UK Biobank Assessment Center at which the prevalence of strabismus was less than 1% were also excluded. From amongst the remaining

participants, there were n=1,022 who self-reported that the reason they wore glasses was because of strabismus. Of these 1,022 participants, the maximum sized set of unrelated participants was selected using the method of Bycroft et al. (2018). This resulted in a sample of n=1,020 unrelated participants with a self-reported history of strabismus ("cases"). There were n=53,052 participants who met all the inclusion criteria but who did not self-report a history of strabismus; these individuals were classified as potential "controls". After excluding individuals from amongst the 53,052 potential controls who were related to one of the cases, and then finding the maximum sized set of unrelated participants, there were n=50,474 potential controls who were unrelated to each other and unrelated to the cases.

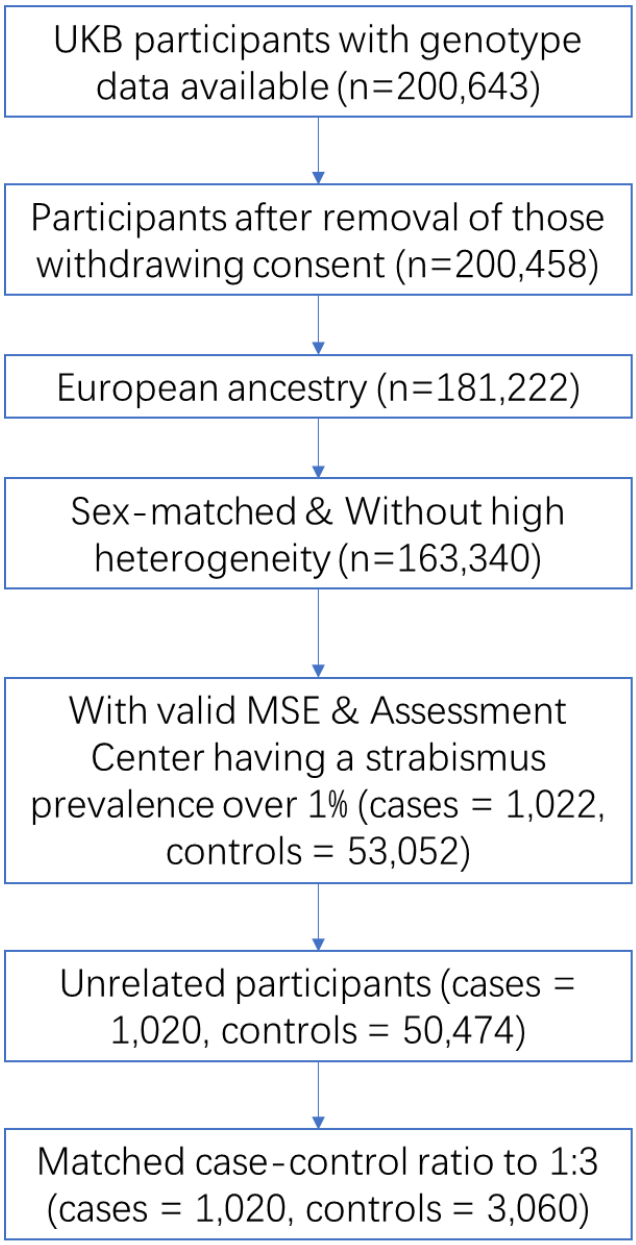


Fig. 3.1 Flow diagram illustrating the selection of UK Biobank participants for the genetic analysis of strabismus sample.

An alternative method to control the type 1 error rate is to set a specific false discovery rate (FDR) for the association results. A disadvantage of the Bonferroni method is that it is a relatively conservative method, which may limit the power of the discovery of true positive results. The FDR method is less conservative. The aim of the FDR approach is to achieve the smallest possible fraction of false signals among all those that are declared to be true. The total number of rejections of the null hypothesis includes both the number of false positive (FP) and true positive (TP). The formula for FDR is FP/(FP+TP). In the controlling procedure, methods for rejecting the null hypothesis were established, such as the Benjamini–Hochberg procedure and Benjamini–Yekutieli procedure, to control the FDR at level alpha. There are precedents for using the FDR approach in GWAS research projects (Nelson et al. 2017; He et al. 2021).

Four criteria were applied for the selection of control participants: 1) cases and controls attended the same Assessment Center; 2) cases and controls had the same sex; 3) cases and controls had the same year of birth; and, 4) cases and controls had the same age (measured in whole years). For each individual in the case group, if three or more individuals met all those four criteria, then 3 controls were chosen at random from the matched set; if fewer than 3 individuals met all 4 matching criteria, then 3 controls were chosen at random from the set matching the first 3 criteria. Finally, 1020 cases and 3060 matched control individuals were selected for the GWAS for self-reported strabismus (Fig. 3.1).

The rationale for the choice of case-control ratio was the result of a trade-off between statistical power and computational efficiency. First, the maximum number of cases was chosen, i.e. the number of individuals who met the inclusion criteria and that passed the quality control filters. Then, the number of controls was chosen. A large number of control participants per case participant increases the statistical

power of an analysis. However, the increase in statistical power rapidly plateaus as the case-control ratio changes and, in fact, heavily unbalanced case-control ratios can generate statistical artefacts (Zhou et al. 2018). WES data requires a very large amount of computational storage space and genotyping STRs using WES data is computationally intensive. Therefore, from a consideration of computational resources versus statistical power, the sample size of the control group was chosen as 3x the sample size of the case group, i.e. a case-control ratio to 1:3. Each person in the case group was matched with 3 individuals from the pool of 50,474 potential controls.

3.2.2. Selection of STRs with valid genotype information

The Genome Reference Consortium Human Build 38 reference panel includes 58,887 STR loci on chromosome 17. A two-step process was developed to identify STRs located within the sequenced regions in the UK Biobank WES dataset. In the first step, a series of HipSTR analyses was used to evaluate all 58,887 candidate STRs in a subsample of n=200 randomly selected participants from the full sample of 4,080 individuals. Candidate STRs were split into 589 groups, with each group containing 100 STRs (except for only 87 STRs in the last group) to facilitate parallel computation.

No pre-processing steps were performed on the UK Biobank WES data before running the HipSTR software. However, the first step of the two-step HipSTR genotyping process would have provided a basic level of quality control (QC). In this step, the minimum and maximum reads-per-sample were set at empirically-determined optimum values to eliminate STRs that could not be reliably genotyped. From amongst the total of 58,887 STRs, the first step identified 1,220 (2.07%) STRs that could be successfully genotyped using WES data from UK Biobank participants.

In the second step, a HipSTR analysis was performed to genotype these 1,220 STRs in the full sample of 4,080 participants. Since a relatively low number of STRs were genotyped in the second step, this step was not parallelized.

The HipSTR program (Willems et al. 2017) was used for STR genotyping. This software implements an expectation–maximization (EM) algorithm to genotype STR loci in regions where they have previously been identified. The EM algorithm is used to find the maximum likelihood parameters of a statistical model when the equations cannot be solved directly. In general, these models are used in cases where latent variables exist with unknown parameters and known observations. With a likelihood function that involves all the variables, the maximum likelihood estimate of the unknown parameters is determined by maximizing the marginal likelihood of the observed data.

To remove the PCR stutter artifacts of the incorrect copies of an STR's motif, HipSTR constructs a stutter model θ_x for each STR locus x, which quantifies the probability that PCR stutter adds or removes repeats from the true allele in an observed read, and parameter ρ_s that evaluate the extent of stutter-induced changes (Willems et al. 2017). The sizes of the STR observed in each read for all individuals in the population are used as the 'observations'; the EM algorithm is then applied to estimate the parameters θ_x and ρ_s (Willems et al. 2017). After the parameter estimation step is completed, HipSTR iteratively computes the maximum-likelihood genotypes for each sample and realigns every read relative to the most probable allele. If the same sequence is observed in a sample in two or more alignment runs with stutter artifacts, HipSTR selects the sequence as a new candidate allele (Willems et al. 2017).

Genetic data for the 4080 UK Biobank participants was input to HipSTR in CRAM format (--bam-files command in HipSTR), alongside a GRCh38 reference assembly (--

fasta command in HipSTR). Specifically, I used the GRCh38_full_analysis_set_plus_decoy_hla.fa reference assembly, obtained from the DNAnexus website. In order to reduce the chance of genotyping errors, a minimum sequence read-coverage threshold was set for each STR. Since 200 participants were analysed in the first step of HipSTR analyses used to screen the STRs, STRs were required to have a minimum of 1,000 reads (equivalent to 5 reads per participant) and a maximum of 90,000 (equivalent to 450 reads per participant). In the second step used to genotype the full sample of 4,080 participants, since the number of samples was larger, the variation in STR coverage was expected to be larger, too. Therefore, in order to genotype the majority of participants, the minimum read threshold was retained at 1,000 reads (equivalent to 0.25 reads per participant) and the maximum of 4,000,000 (equivalent to 980 reads per participant).

3.2.3. Chromosome 17 Association Study for Strabismus

A genetic association study for self-reported strabismus was carried out in the discovery sample (1,020 cases and 3,060 controls). A total of 742 STRs on chromosome 17 were tested for association with strabismus using logistic regression (only 742 STRs out of the total of 1220 STRs were studied, since the remainder of the STRs exhibited no variation in allele length in this cohort, i.e. they were monomorphic). The average length of the two alleles of each STR genotype was used as a predictor variable. Age, sex, refractive error averaged between the two eyes, and the first 10 ancestry principal components were included as covariates. Logistic regression models were fitted using the glm function in R. This approach of using the average length of an STR in a regression analysis has previously been adopted by Fotsing et al. (2019).

As an alternative to logistic regression, a Chi-squared test was performed to test for association between each of the 742 STRs and strabismus. A 2 x 3 table was created for each STR (Table 3.1). The two rows of the table were used to record genotype counts in cases and controls, respectively. The three columns of the table were used to record the count of alleles of different length. As an example, consider an STR with motif 'TC' and alleles of length 10, 12, 14, 16 and 18. The middle column recorded the count of alleles with length=12 in cases and in controls. The upper column recorded the count of alleles shorter than 12 in cases and controls. The lower column recorded the count of alleles longer than 12 in cases and controls. The distribution of allele lengths in cases versus controls was tested with a Chi-squared test (except if one or more cell of the table had a count less than 5, when a Fisher's exact test was implemented instead). The process of creating a 2 x 3 table and testing for association with strabismus was then repeated based on counts of alleles with length below, equal to, or above 14, and then again based on counts of alleles with length below, equal to, or above 16, etc. For each STR, the number of statistical tests was counted (e.g. 3 Chi-squared tests would have been required to test the STR with allele lengths 10, 12, 14, 16 and 18). The lowest p-value amongst these tests was chosen to represent the p-value of this STR. To avoid the increase of false positive rate through multiple comparisons, a Bonferroni correction was applied for each STR (e.g. requiring the p-value to be lower than 0.05/3=0.0167 for the STR with allele lengths 10, 12, 14, 16 and 18). This approach of testing STRs for association with a phenotype has previously been reported by (Pritchard and Rosenberg 1999). Statistical software used to perform GWASs in this study include Plink2.0 and R x64 4.0.3 (R Core Team (2020) ; Chang et al. 2015). Logistic regression, Chi-squared test or Fisher's exact test are the main statistical methods used by these software packages.

A Manhattan plot is applied for the visualization of the GWAS results. Each data point

in a Manhattan plot corresponds to the statistical test result (negative $\log_{10}$ of the p-value) of one genetic variant. The genomic coordinates are displayed on the x-axis, with the negative logarithm of the association *p*-value for each genetic variant displayed on the y-axis.

Table 3.1 Contingency table of the count of A alleles and non-A alleles in cases and controls samples

Alleles	The Number of			Total No. of Alleles
	A-	A[a]	A+	
Cases	m1[b]	mc - m1 - m2	m2	mc
Controls	n1	nc - n1 - n2	n2	nc
Total No. of Alleles	m1+n1	mc+nc-(m1+n1+m2+n2)	m2+n2	mc+nc

[a] A^- denotes non-A alleles that have a shorter average length than A alleles; A^+ denotes non-A alleles that have a longer average length than A alleles.

[b] *m1,m2* denote the number of A^- and A^+ alleles in the cases, respectively; and *n1, n2* denote the number of A^- and A^+ alleles in the controls respectively. *mc* and *nc* denote the number of alleles in cases and controls.

3.2.4. Post-association Study Analyses

A strong association with self-reported strabismus in UK Biobank participants was reported previously for SNP rs75078292 on chromosome 17 (Plotnikov et al. 2019). Therefore, a conditional association test was carried out to determine if the association between STR Human_STR_613083, which was the only STR with pairwise LD ($r^2 > 0.1$) with rs75078292, was independently associated with strabismus. For this test, a conditional logistic regression analysis was conducted, with strabismus as the outcome, Human_STR_613083 genotype as the primary independent variable, and with SNP rs75078292 genotype included as an additional covariate along with age, sex, refractive error averaged between the two eyes, and the first 10 ancestry principal components.

3.3. Results

3.3.1. Validation of Self-reported Strabismus in UK Biobank Cohort

Among the 4,080 case and control participants, the cases with self-reported strabismus had an 11.1-fold higher prevalence of self-reported unilateral amblyopia, a 2.4-fold higher prevalence of 1.00 D or more anisometropia, and a more hypermetropic refractive error (median +2.36 vs. +0.24 D) compared to controls. Moreover, individuals in the case group had a much higher prevalence of early age-of-onset of wearing glasses (age started wearing glasses $\leqslant$7 years): 73.0% vs. 4.6% in cases vs. controls). Furthermore, a 2.2-fold lower proportion of cases had a high level of visual acuity (visual acuity $\leqslant$ 0.0 logMAR in both eyes): 20.6% vs. 45.6% (Table 3.2; Fig. 3.2). The sex and age of cases and controls was well-matched (by design), and

their Townsend Deprivation Index, a measurement of socioeconomic status, did not show a significant difference between cases and controls (-1.91 vs. -2.09; $p = 0.54$).

Table 3.2 Demographic and ocular characteristics of the UK Biobank strabismus case-control sample

Variable		Total (n=4,080)	Cases (n=1,020)	Controls (n=3,060)	*p* value
Female	N (%)	2,528 (62.0%)	632 (62.0%)	1,896 (62.0%)	1
Self-reported unilateral amblyopia	N (%)	469 (11.5%)	374 (36.7%)	95 (3.1%)	<1.0E-99
Both eyes VA ⩽0.0 logMAR	N (%)	1,065 (39.33%)	210 (20.6%)	1,395 (45.6%)	7.9E-44
VA difference ⩾0.2 logMAR	N (%)	1,034 (25.3%)	456 (44.7%)	578 (18.9%)	3.89E-65
Better VA⩽0.0; VA difference⩾ 0.2	N (%)	727 (17.8%)	326 (32.0%)	401 (13.1%)	2.78E-45
Anisometropia⩾ 1.00 D	N (%)	940 (23.0%)	421 (41.3%)	519 (17.0%)	2.06E-57
Anisometropia⩾ 2.00 D	N (%)	374 (9.2%)	200 (19.6%)	174 (5.7%)	1.28E-40
Age (years)	Median (IQR)	59.75 (53.00 to 64.33)	59.75 (53.06 to 64.35)	59.75 (52.92 to 64.33)	9.25E-02
Refractive error (D) average of 2 eyes	Median (IQR)	+0.48 (-0.74 to 1.82)	+2.36 (0.44 to 4.45)	+0.24 (-1.04 to 1.14)	<2.2E-16
Anisometropia (D)	Median (IQR)	0.41 (0.17 to 0.92)	0.77 (0.29 to 1.67)	0.34 (0.15 to 0.73)	<2.2E-16
Age started wearing glasses (years)	Median (IQR)	21.00 (8.00 to 45.00)	5.00 (3.00 to 8.00)	40.00 (16.00 to 48.00)	<2.2E-16
Townsend Deprivation Index	Median (IQR)	-2.06 (-3.55 to 0.50)	-1.91 (-3.54 to 0.52)	-2.09 (-3.56 to 0.50)	5.42E-01

Abbreviation: IQR = interquartile range.

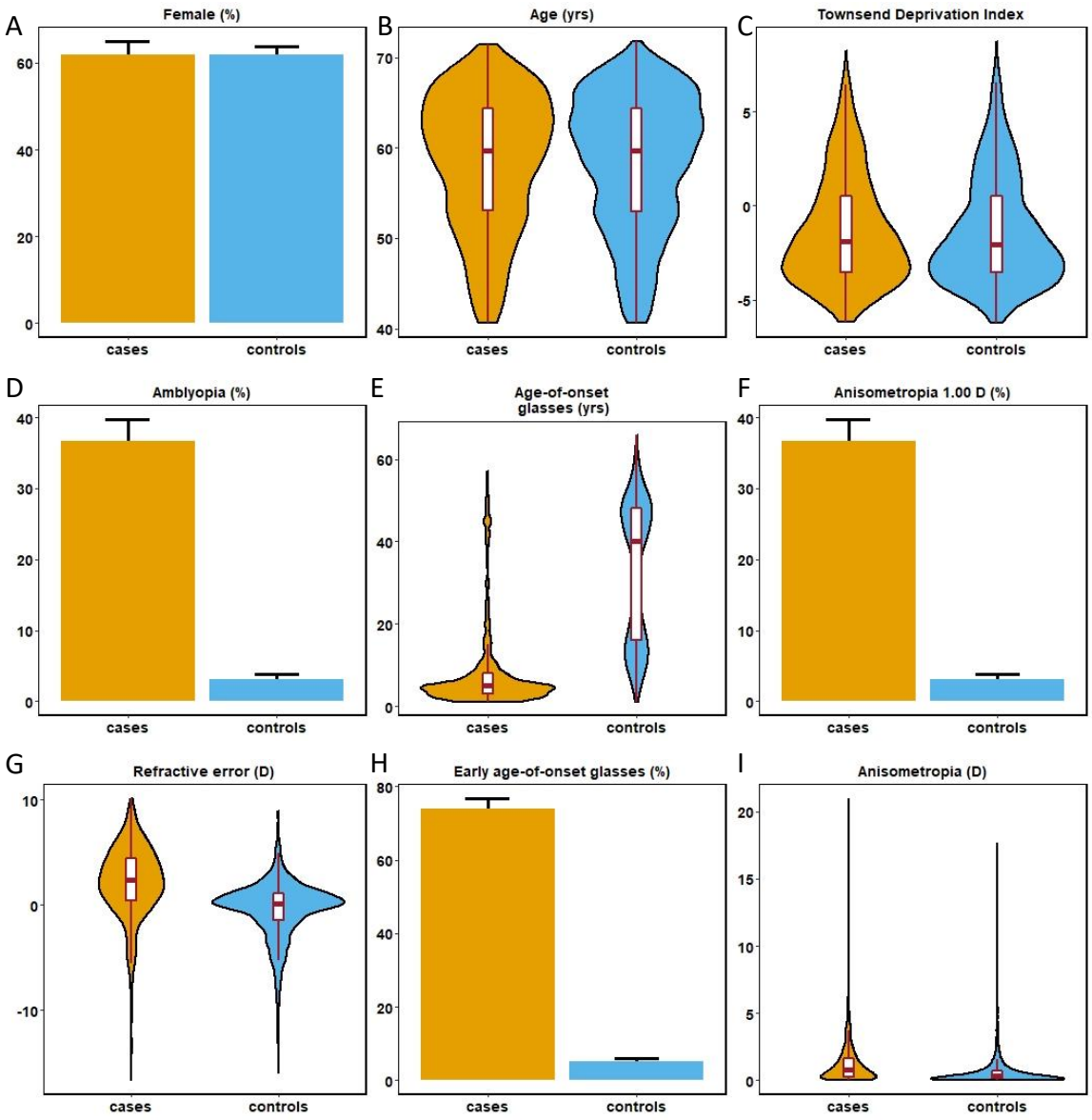


Fig. 3.2 Demography and clinical characteristics of cases and controls (n=1,020 cases and n=3,060 controls). (A) Bar chart for the percentage females. (B) Violin plot for the age of participants. (C) Violin plot for Townsend Deprivation Index of participants. (C) Violin plot for the age-of-onset wearing glasses. (D) Bar chart for the percentage of amblyopia. (E) Violin plot for age-of-onset of glasses. (F) Bar chart for the percentage of anisometropia above 1.0 D. (G) Violin plot for refractive error (H) Bar chart for the percentage of early age-of-onset of glasses. (I) Violin plot for anisometropia of participants. Early age-of-onset of glasses was defined as $\leqslant$7 years. Bar chart error bars denote 95% confidence interval. Box plots are superimposed over the violin plots (the thick horizontal line corresponds to the median, the white rectangular box

corresponds to the interquatile range, and the end of upper and lower whiskers to the largest and smallest sample within 1.5 times the interquatile range).

3.3.2. Call Rate Assessments

To evaluate the reliability of the genotype information obtained from HiPSTR, the call rate of each sample and the call rate of each STR were examined. The call rate of a sample is defined as the percentage of called STRs for which the genotype value is not null, as a proportion of the total number of STRs in the dataset. Similarly, the call rate of an STR is defined as the percentage of called samples for that STR which are not null, as a proportion of the total number of samples in the dataset. The typical value to exclude the genetic variants with a low call rate is <95% (Anderson et al. 2010).

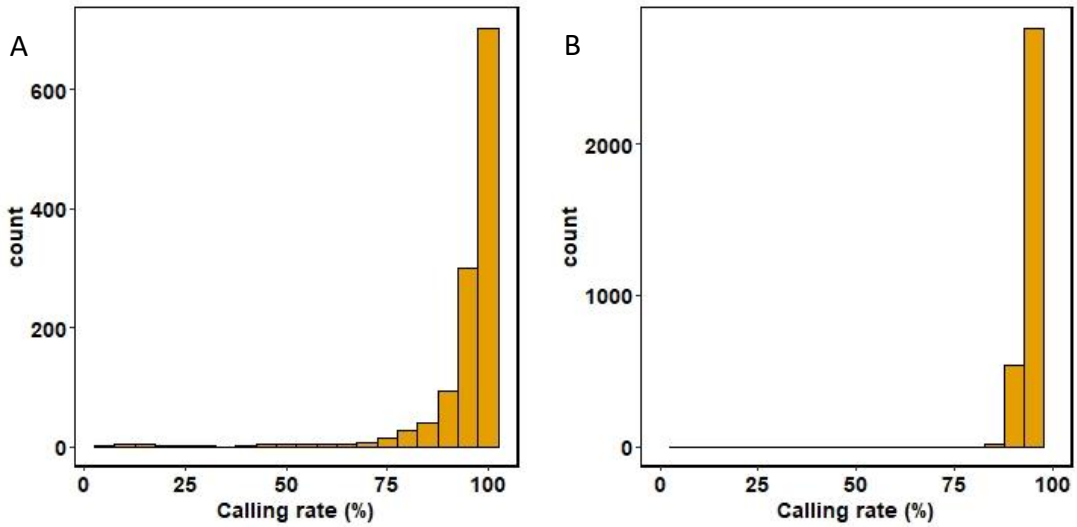


Fig. 3.3 Histogram of call rates of STRs and samples. (A) Distribution of call rates of 1220 STRs. (B) Distribution of call rates of 4080 samples after the exclusion of STRs with a call rate <95%.

A high STR call rate refers to the situation in which a high proportion of samples can be genotyped reliably. In general, a high call rate is indicative of high quality (i.e. accurate) genotype data. However, in practice, calling genotypes with high certainty does not necessarily imply high quality genotype data, because failure to call may be dependent on genotype. For example, rare homozygous genotypes may on average, have lower probabilities, thus introducing bias to allele frequencies based only on genotypes (Clayton et al. 2005). Furthermore, a high calling threshold could unnecessarily exclude missing genotypes, which may result in reduced genomic coverage. The call rate threshold has a large impact on the quality of the genotype data. If it is set too low, erroneous genotypes can be assigned. If it is set too high, then information from a large number of genetic markers may be wasted. Classically, markers with a call rate less than 95% are removed from further study (Fisher et al. 2008; Silverberg et al. 2009).

Compared to the call rate of samples, the call rate of STRs displayed a relatively greater variation. Among the full set of 1220 genetic variants, 908 STRs had a call rate above 95%, while 100 STRs had a call rate lower than 90% (Fig. 3.3A). STRs with a call rate lower than 95% were excluded from further analysis, in order to avoid STRs that were challenging to genotype and thus potentially more susceptible to genotype call errors.

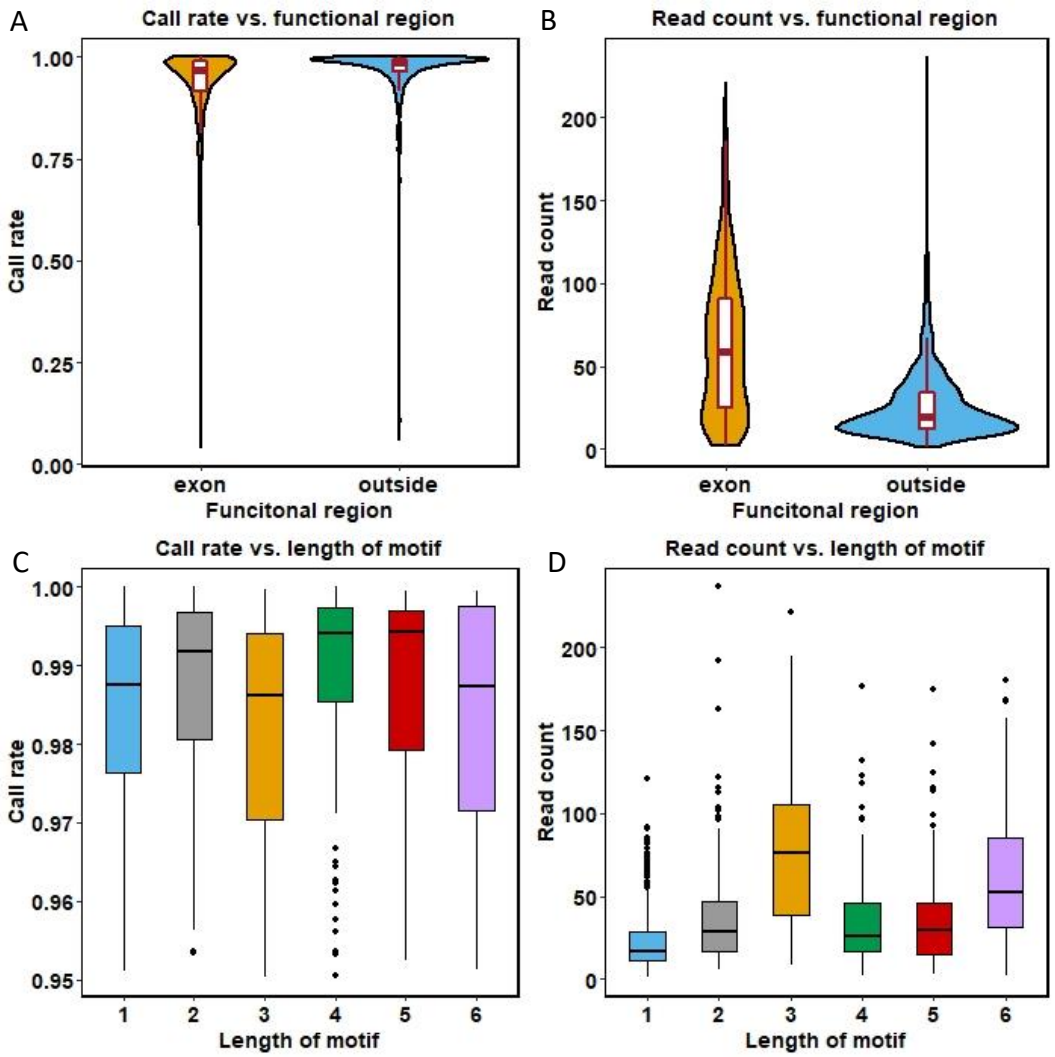
After the exclusion of STRs with a call rate <95%, the call rates of 4080 participants were examined for the remaining set of 908 STRs. The call rate of samples exhibited lower variation ($F = 39.91$, $p = 2.2E\text{-}16$; Fig. 3.3B) but a similar average level to the call rates of STRs ($t = 0.32$, $p = 0.75$). 2282 samples had a call rate greater than 95%, and all 4080 samples had a call rate greater than 85%. To keep the number of cases and controls constant at the planned ratio of 1:3, all the samples were included in the further study.

Further analysis revealed that the 1220 STRs were located both inside and outside exon regions. 444 STRs were found within exons and 776 were situated outside exon regions. Read coverage was also considered, i.e. how many reads were obtained for each STR, on average. The average call rate of STRs in exon regions was lower than the call rate of STRs located outside exons: inside vs. outside mean call rate = 0.93 vs. 0.95, respectively, p = 1.2E-03; Fig. 3.4A. The mean read count for STRs within exon regions was significantly higher than for those outside exon regions: inside vs. outside mean read count = 64.1 vs. 27.2, p = 2.2E-16 (Fig. 3.4B), suggesting the genotype results of STRs in the exon regions may be more reliable. Shapiro-Wilk's test revealed the distribution of call rates was non-normal (p < 2.2E-16); therefore, the correlation between call rates and read counts was estimated with Spearman's rank correlation coefficient. This revealed a negative correlation between ranks of call rates and ranks of read counts (ρ = -0.16, p = 3.67E-08; Fig. 3.4E), but the fact that the correlation coefficient was close to zero indicated that the correlation trend was not very strong.

The relationship between the call rate and other STR features was also studied. Since the distribution of call rates was non-normal, the mean call rate of each group, which was classified by the length of motif, was compared using the Kruskal-Wallis rank sum test. A significant difference in the mean call rate was found (p = 9.51E-13; Fig. 3.4C). The 6-bp motif group had a lowest call rate (92.7%), while the 2-bp motif group had the highest call rate (97.0%). The 6-bp motif group had a significantly different mean call rate compared to the 2-bp and 4-bp motif groups (Wilcoxon rank sum test $p_{2,6}$ = 4.50E-04, $p_{4,6}$ = 9.31E-03). Meanwhile, the mean read count per sample for each group, which was classified by the length of motif, was also compared using the Kruskal-Wallis rank sum test. A significant difference in the mean call rate was found (p < 2.2E-16; Fig. 3.4D). The 3-bp motif group had the highest

read count (76.7 reads), and the 1-bp motif group had the lowest read count (22.9 reads).

The correlation between call rate and motif length was estimated with Spearman's rank correlation test. This revealed a negative correlation between ranks of call rate and motif length (ρ = -0.09, p = 1.27E-03; Fig. 3.4F). However, the coefficient was weak. Meanwhile, a stronger negative correlation was found between read count and motif length (ρ = -0.26, p < 2.2E-16).


A
Call rate vs. functional region
Call rate
1.00
0.75
0.50
0.25
0.00
exon
outside
Funcitonal region
B
Read count vs. functional region
Read count
200
150
100
50
0
exon
outside
Funcitonal region
C
Call rate vs. length of motif
Call rate
1.00
0.99
0.98
0.97
0.96
0.95
1
2
3
4
5
6
Length of motif
D
Read count vs. length of motif
Read count
200
150
100
50
0
1
2
3
4
5
6
Length of motif

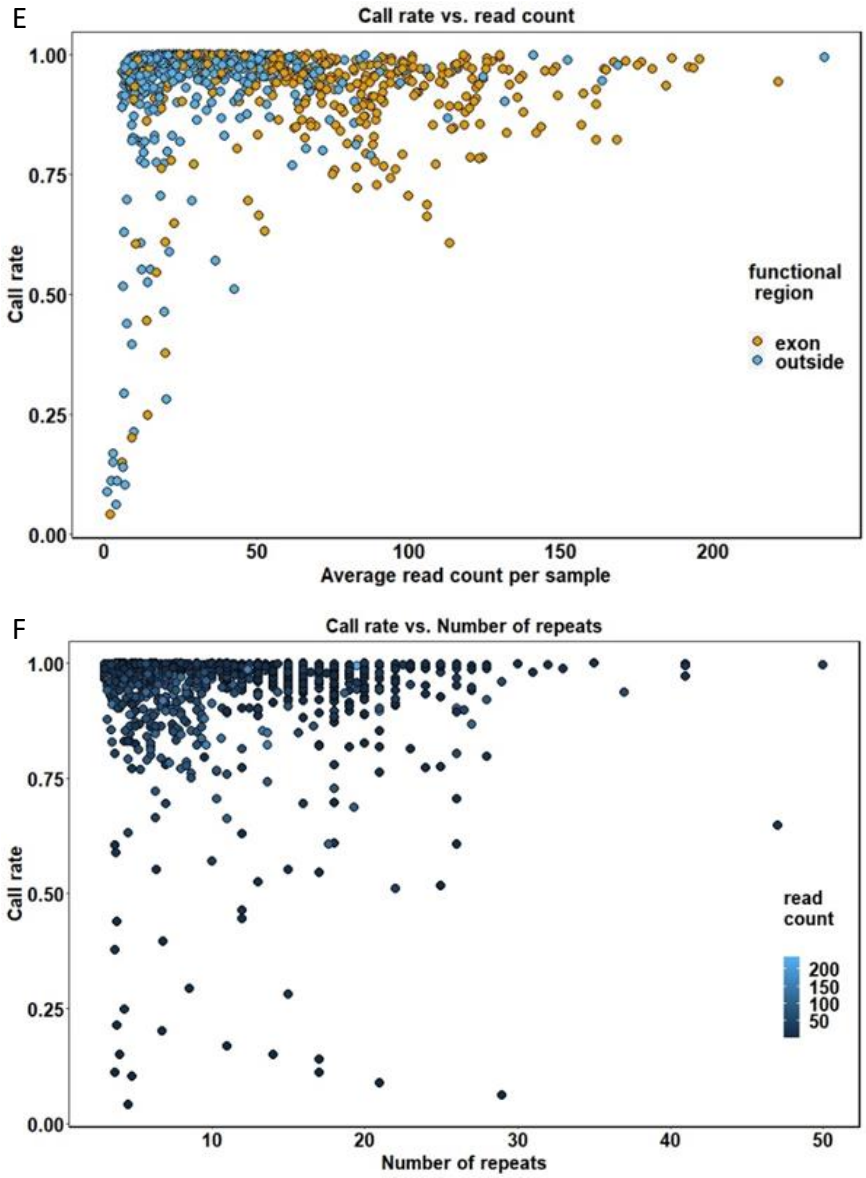


Fig. 3.4 The quality of the genotype information for 1220 STRs on chromosome 17

(A) The distribution of call rate versus functional region. (B) The distribution of read count versus functional region. (C) Graph of call rate versus motif length. Box plots are superimposed over the violin plots (the thick horizontal line corresponds to the median, the white rectangular box corresponds to the interquatile range, and the end of upper and lower whiskers to the largest and smallest sample within 1.5 times the interquatile range). (D) Graph of read count versus motif length. (E) The relationship between call rate versus read count. Different colors represent different functional regions. (F) The relationship between call rate versus motif length. The shading denotes the average read count.

3.3.3. Mapping of STRs on Chromosome 17

To study the association between a phenotype and STR genotypes, the STRs must be polymorphic. 908 of the 1220 genotyped STRs had a call rate above 95%. Surveying the number of alleles of each STR revealed that 742 out of the 908 STRs were polymorphic. The 166 monomorphic STRs were excluded from further analysis.

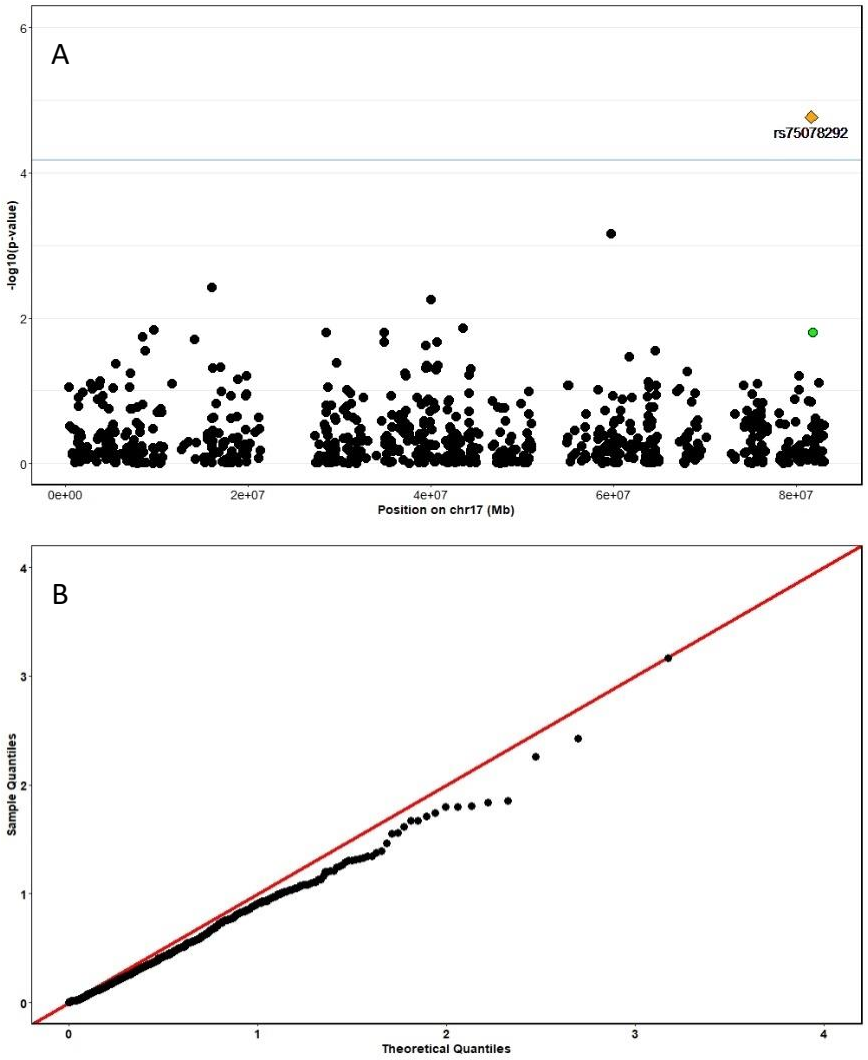


Fig. 3.5 Manhattan plot and quantile-quantile (Q-Q) plot for p-value of logistic regression analysis. (A) Manhattan plot for association study, the yellow diamond denotes the result of the lead SNP rs75078292, the green symbol denotes the single STR that had a squared-correlation >0.1 with the lead SNP. The light blue line corresponds to a *p*- value of 5.74E-05. (B) Q-Q plot for *p*-value of STRs, in which x-coordinate denotes the theoretical quantile for a uniform distribution, and y-coordinate denotes the observational sample quantile distribution (B).

A logistic regression analysis testing for association of self-reported strabismus case/control status and the average length of the 2 STR alleles carried by each participant was carried out for the 742 polymorphic STRs on chromosome 17. This analysis identified no variant that was significantly associated with the trait after correction for multiple testing. Applying the Bonferroni method, the threshold *p*-value for declaring statistical significance was set as 0.05/742 = 5.74E-05. Of the 742 variants, the most strongly associated STR was Human_STR_596782: OR=0.33, 95% CI 0.17-0.62, p = 6.88E-04, Fig. 3.5A. The 6 variants with the lowest *p*-values are listed in Table 3.2. None of the variants had a p-value below the threshold of p < 5.74E-05.

In order to check for systematic bias due to population stratification, the genomic inflation factor (λ_{GC}) was calculated. There was no evidence of population stratification ($\lambda_{GC} = 0.65$, Fig. 3.5B). In the Q-Q plot, the observed p-value distribution showed a distribution close to theoretical uniform distribution, suggesting no excess of STRs strongly associated with the strabismus phenotype.

Table 3.3 Lead variants for attaining relatively lower p-values in logistic regression for chromosome 17

STR	Start Position (bp)	Reference Repeat Number	Repeat Motif	OR	95% C.I.	p-value
Human_STR_596782	59685042	11	T	0.33	0.17, 0.62	6.88E-04
Human_STR_568152	16071266	12	A	0.66	0.49, 0.87	3.73E-03
Human_STR_581897	39996121	15	A	1.24	1.06, 1.44	5.55E-03
Human_STR_584893	43496098	35	A	0.33	0.13, 0.79	1.39E-02
Human_STR_564072	9686927	7.5	GT	0.62	0.42, 0.90	1.46E-02
Human_STR_613083	81719696	4.2	ACACCC	0.94	0.90, 0.99	1.55E-02

Abbreviations: OR = odds ratio; 95% C.I. = 95% confidence interval of odds ratio.

A Chi-squared test was also used to test for an association between self-reported strabismus case/control status and STR genotype, for the 742 polymorphic STRs on chromosome 17. The Chi-squared test analysis identified 1 variant significantly associated with the trait. The threshold p-value for declaring statistical significance was again set at 0.05/742 = 5.74E-05. The most strongly associated STR variant was Human_STR_584893 (p = 3.50E-06; Fig. 3.6A). The lead variants with lowest p-value were listed in Table 3.3. No other variant was statistically significant (all p > 5.74E-05).

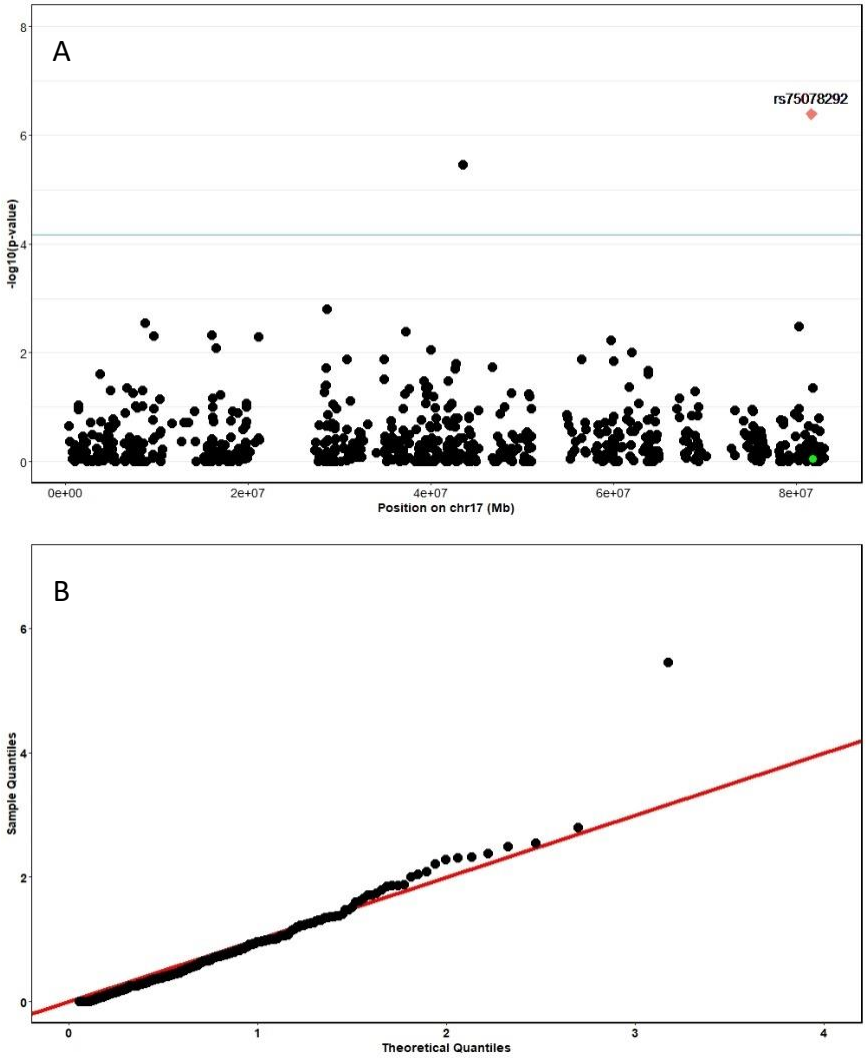


Fig. 3.6 Manhattan plot and Q-Q plot for the Chi-squared test analysis (A) Manhattan plot for association with self-reported strabimsus. The orange diamond denotes the result of the lead SNP, rs75078292. The the green symbol denotes the only STR with a squared correlation coefficient >0.1 with the lead SNP. The light blue line corresponds to the *p* -value threshold of 5.74E-05. (B) Q-Q plot for *p*-value of STRs, in which x-coordinate denotes the theoretical quantile in uniform distribution, and y-coordinate denotes the observational sample quantile distribution.

Table 3.4 Lead variants attaining relatively low p-values in Chi-squared test for association with self-reported strabismus for STRs on chromosome 17

STR	Start Position (bp)	Reference Repeat Number	Repeat Motif	p - value
Human_STR_584893	43496098	35	A	3.50E-06
Human_STR_573636	28632840	4.2	CCCCT	1.58E-03
Human_STR_563295	8735291	19	GT	2.79E-03
Human_STR_612241	80290809	17	T	3.23E-03
Human_STR_579636	37270866	20	A	4.11E-03
Human_STR_568152	16071266	12	A	4.67E-03

In the Q-Q plot, the result of genomic inflation factor was low ($\lambda_{GC} = 0.40$, Fig. 3.6B). The observed p-value distribution was slightly enriched for moderately low p-values, in the region of minus log(P) = 2 to 3, which may have indicated either true enrichment or spurious association. The trend in the Q-Q plot for the Chi-squared analysis was not consistent with the trend of the Q-Q plot for the logistic regression analysis (Fig. 3.5B).

To explore whether the inclusion of covariates in the logistic regression analysis, but not in the Chi-squared analysis, explained the difference in results, the logistic regression analysis was repeated using the same genotype coding scheme adopted for the Chi-squared test. This revealed that the absence of covariates was responsible for the different results obtained with the two tests, and suggested the very low p-value for Human_STR_584893 (p = 3.50E-06) was a spurious finding (Table 3.5).

Table 3.5 Logistic regression results for Human_STR_584893 and strabismus.

Condition	OR	95% C.I.	p-value
With no covariates	0.21	0.091,0.50	3.47E-04
With covariates	0.33	0.13,0.79	1.39E-02

Abbreviations: OR = odds ratio; 95% C.I. = 95% confidence interval of odds ratio.

3.3.4. Correlation Study with Single Nucleotide Polymorphism

Human_STR_613083 was the only STR for which the average STR allele length showed a moderate level of linkage disequilibrium ($r^2 > 0.1$) with the genotype of lead SNP rs75078292. In order to identify if Human_STR_613083 had an independent association with strabismus, a conditional analysis was conducted in which the association of the STR was tested while the genotype of SNP rs75078292 was included as a covariate. The *p* value for Human_STR_613083 was increased in the conditional analysis compared to the original analysis (*p* = 7.40E-01 vs. 1.55E-02). Furthermore, the *p* value for rs75078292 remained low in both the conditional and unconditional analyses (*p* = 4.81E-04 vs. 1.71E-05). Thus, the conditional analysis suggested that the only STR in LD with rs75078292 did not have an independent association with self-reported strabismus.

3.4. Discussion

This study screened exonic STRs on chromosome 17, but did not identify any genetic variants showing significant association with self-reported strabismus. The participants in the case and control groups all had European ancestry and genomic inflation factors were computed to assess whether or not population stratification was present. Rigorous quality control metrics were applied to ensure the allele frequency differences reflected the genuine locus-specific associations rather than population stratification.

The method used in this study was justified for the purpose of a case-control association study to identify genetic variants associated with a specific trait. First, the trait must be heritable. Previous studies provide support for the hypothesis that

there are genetic effects contributing to susceptibility for strabismus (Maconachie et al. 2013; Ye et al. 2014; Shaaban et al. 2018; Plotnikov et al. 2019). Second, the extent of misclassification bias should be less than 5% (The Wellcome Trust Case Control Consortium 2007). The statistic power could be affected if a substantial proportion of the controls meet the criteria for inclusion as a case. Since the known prevalence of strabismus in the white British population is about 2-4%, participants selected as controls who did not self-report strabismus are considered effective to meet the second criteria. Third, the proportion of true positives in the case group should be sufficient. We analyzed the level of comorbid factors such as anisometropia, self-reported amblyopia, and refractive error in self-reported strabismus case and control group. The groups featured significantly different characteristics, as expected for a set of strabismus cases and controls.

The analysis involving average spherical equivalent and other covariates reduced the association of Human_STR_584893 compared to the univariate analysis. This suggested the originally observed association of Human_STR_584893 with strabismus occurred due to confounding.

When examining the association of STR genotype and strabismus, it was noticed that three of the covariates included in the logistic regression analysis showed a significant association with strabismus. These were the average spherical equivalent ($p < 2E-16$), age ($p = 3.88E-05$), and principal component 1 ($p = 4.61E-03$).

The association between refractive error and strabismus is consistent with previous publications. A cross-sectional study involving 4273 children aged 6-8 years old from Hong Kong revealed an association of strabismus with myopia and hyperopia (Zhang et al. 2021), while one Korean population-based study found an increased prevalence of exotropia among people with severe myopia, while esotropia was significantly associated with hyperopia (Lee and Mackey 2021). A meta-analysis involving 23,541

study subjects identified a dose-related effect for hyperopia as a risk factor for concomitant esotropia (Tang et al. 2016). In the Tang et al. study, a hyperopic refractive error showed an association with an increased likelihood for self-reported strabismus (OR = 3.00, 95% CI = 2.71-3.37, p < 2E-16).

In the analysis model that included all of the covariates, age was negatively associated with self-reported strabismus in the current study (OR = 0.84, 95% CI = 0.80-0.90). However, when the other covariates were removed, the association of age and strabismus was no longer significant (p = 0.981). Furthermore, the mean and variance of age were not significantly different between cases and controls (t-test p = 0.981; F-test p = 0.978). Therefore, the association between age and strabismus in the model including all covariates could either have been due to a true association or due to confounding.

The ancestry variable PC1 was found to be positively associated with self-reported strabismus (OR = 1.14, 95% CI = 1.07-1.22, p = 4.61E-03). This association was minimally affected if assessment center was included as an additional covariate in the analysis (OR = 1.14, 95% CI = 1.04-1.26; p = 5.29E-03), suggesting that differences relating to assessment centers were not the reason for the association of PC1 with strabismus. Furthermore, an analysis that included “northing” and “easting” coordinates, corresponding to each participant place of birth, did not appreciably attenuate the association of PC1 with strabismus (OR = 1.14; p = 9.77E-03). PC1 differentiates the ethnic backgrounds of individuals, and natural selection pressure has a strong effect on genetically distinct subgroupings across the world. The lead genetic variant rs75078292 at the *NPLOC4-TSPAN10-PDE6G* locus of chromosome 17 is associated with hair and skin pigmentation (Wollstein et al. 2017). Since pigmentation traits have been under strong selection pressure during human evolution, rs75078292 was also included as a covariate in the analysis; however, the association of PC1 still minimally affected (OR = 1.14, 95% CI = 1.04-1.26, p = 5.11E-

03). The association of PC1 with strabismus was robust, but the reason for the association could not be identified.

Lambda-GC quantifies the inflation in the test statistics from non-associated markers, relative to the expectation under the null hypothesis for either the median (Devlin and Roeder 1999), or mean (Reich and Goldstein 2001) of the chi-squared distribution. The factor '0.4549' is the median and mean of the one-degree-of-freedom chi-squared distribution. Therefore, the lambda-GC inflation factor is calculated as the median of the observed chi-squared statistics divided by 0.4549 (Devlin and Roeder 1999). Previous studies revealed high variability of lambda-GC when few SNPs are genotyped, while its variability decreases substantially at higher numbers of SNPs (Dadd et al. 2009). The small sample property was found to increase the type 1 error of testing for association, while the anti-conservative and conservative behaviour depended on parameters such as sample size imbalance between groups and the presence of population stratification (Dadd et al. 2009). A more conservative pattern was frequently identified under conditions with lower case-control sample size mismatch and an absence of population stratification, especially when relatively few SNPs were included in the GWAS (Dadd et al. 2009). In my study, the low lambda-GC may have resulted from high variability of the test due to the low number of the STRs included in my analysis, along with linkage disequilibrium between variants. The sample selection criteria can impact the case-control sample size mismatch and the level of population stratification. Therefore, the conservative lambda-GC value in my study for the association of self-reported strabismus with genotyped STRs is consistent with a low level of population stratification.

A previous study in UK Biobank found the *NPLOC4-TSPAN10-PDE6G* gene cluster was associated with the risk of strabismus, and the finding was replicated in a sample of 7-year-old children with clinician-diagnosed strabismus (Plotnikov et al. 2019); 20

SNPs within the gene cluster were found in nearly perfect linkage disequilibrium, and two SNPs in TSPAN10, a missense variant rs6420484 and a 4-bp deletion variant rs397693108, were predicted to have functional influence (Plotnikov et al. 2019). Another study involving a large adult population sample from Finland replicated the association of the *NPLOC4-TSPAN10-PDE6G* with both convergent and divergent strabismus (Plotnikov et al. 2022). In the current study, rs75078292, the lead variant in the *NPLOC4-TSPAN10-PDE6G* gene cluster (Plotnikov et al. 2019) was included as a covariate in conditional logistic regression analyses, to analyze the association of STRs with strabismus independently of the known association. The only STR in linkage disequilibrium with rs75078292 was not significantly associated with self-reported strabismus, suggesting that an exonic STR is unlikely to drive the association with strabismus at this locus.

Future studies of additional samples are needed to replicate and extend the existing results, especially samples for which strabismus has been clinically diagnosed during childhood. These future studies should evaluate if STRs on all human chromosomes, not just chromosome 17, are associated with strabismus. These studies should also examine the association in other ethnicities, and confirm their potential inheritance patterns. Studies with a larger sample size are necessary to determine if Human_STR_584893 is associated with strabismus case-control status.

Chapter 4. Genome-wide Association Study for High Myopia

4.1. Introduction

High myopia is an extreme type of myopia, defined as a refractive error of less than -6.00 diopters (D) or an axial length longer than 26mm (Young et al. 1998b). The axial length is the distance from the corneal surface to the retina (Hitzenberger 1991; Schmid et al. 1996). The average axial length for a human emmetropic eye is about 23.5mm (Gordon and Donzis 1985). Eyeballs with more myopic spherical equivalent have longer axial lengths ($r = -0.90$, $p < 0.001$), while the corneal curvature is flatter in eyes with longer axial lengths ($r = -0.22$, $p=0.003$) (Schmid et al. 1996). The average axial length for a human emmetropic eye is about 23.5mm (Gordon and Donzis 1985). Eyeballs with more myopic spherical equivalent had longer axial length ($r = -0.90$, $p < 0.001$) and decreased endothelial density ($r = 0.20$, $p = 0.037$), while the corneal curvature was flatter in eyes with longer axial length ($r = -0.22$, $p=0.003$)(Schmid et al. 1996). The prevalence of myopia is showing an increasing trend all over the world, especially in East Asia, affecting as many as 90% of those of school-leaving age (Vitale et al. 2009; Morgan et al. 2012; Williams et al. 2015). The current trend predicts the number of people affected by myopia will increase from 1.4 billion to 5 billion by 2050, affecting about half of the world's population, and almost 10% of the affected people will get high myopia (Holden et al. 2016). The elongated axial length increases the risk of complications, such as retinal detachment, glaucoma, and myopic macular degeneration. These complications can further cause blindness due to high myopia (Saw et al. 2005; Fujimoto et al. 2010; Verhoeven et al. 2015).

Population-based epidemiological studies have identified a significant influence of genetic factors on myopia onset and progression. The broad-sense heritability measures the proportion of variance in a trait that is attributed to genetic variance. The broad-sense heritability may also capture effects relating to gene-by-gene

interaction or gene-by-environment interaction. However, it is generally not the best predictor of the level of resemblance of offspring and their parents. Instead, the narrow-sense heritability is defined as the fraction of variance in a trait that can be attributed to 'additive' genetic variance. The narrow-sense heritability is used to measure how variation among individuals is influenced by genetic differences that are, on average, passed from parents to offspring. The broad-sense heritability is commonly estimated in twin studies, while narrow-sense heritability can be assessed in family-based samples. Both approaches require simplifying assumptions to be made, such as ignoring assortative mating.

In large-scale twin studies, the heritability of myopia has been estimated to be up to 90% (Hammond et al. 2001; Lyhne et al. 2001). This high heritability estimate in twin studies can be attributed to the limited environmental variation within twin pairs, combined with the method's underlying assumptions, for example, the assumption that twins share a common environment. Studies have also investigated the heritability of high myopia, as a binary phenotype, and GWAS analyses for high myopia have also been reported, as discussed below.

A GWAS for high myopia identified 6 associated loci in an East and Southeast Asian population (Meguro et al. 2020). This study also highlighted the role of the nervous system in its pathogenesis. A meta-analysis study that combined six case-control association studies found that SNP rs644242 in the PAX6 gene had a suggestive association with high myopia (Tang et al. 2014). However, the inheritance pattern of high myopia has a more complex pattern than expected for a monogenic trait. High penetrance autosomal dominant loci were reported to make a contribution to cases of high myopia (Farbrother et al. 2004). A strong association was found between high myopia in parents and the onset of myopia in children, while in siblings there was a weaker association with the level of myopia and no effect on the age-of-onset of myopia (Liang et al. 2004). It may be that high myopia clusters in families to a greater

extent than moderate myopia: the familial aggregation of high myopia can be explained by the autosomal dominant inheritance pattern in several chromosome loci identified in genetic linkage analysis(Young et al. 1998a; Naiglin et al. 2002; Lam et al. 2003; Paluru et al. 2003b). Guggenheim et al. (2000) calculated the recurrence-risk ratio of high myopia in siblings to be $\lambda s = 20$, in a population-based sample of Danish teenagers. Farbrother et al. (2004) estimated $\lambda s = 4.9$ (95% CI: 2.8 – 7.6) based on the presence of high myopia among siblings through a questionnaire, which surveyed the age of onset of spectacle wear of 9.1 years or younger. The inconsistency of heritability among studies of makes it necessary to consider high myopia as a 'complex trait'. In general, linear regression analysis is a more powerful tool for quantitative traits. However, a violation of the assumption of linear relationship between genetic variants and trait affects the accuracy of the estimation. When the linear assumption is incorrect – for example a threshold-related relationship – a case-control logistic regression could potentially offer higher power than a linear regression analysis.

Gene-gene interactions and gene-environment interactions are also implicated in myopia pathogenesis. Although myopia usually exhibits familial aggregation (Lee et al. 2001; Wojciechowski et al. 2005; Fotouhi et al. 2007), genetic factors alone are unable to account for the rapid increase in the prevalence of myopia over the past few decades. Epidemiological studies have identified the environmental effects associated with myopia, such as insufficient outdoor activities, excessive near work, more time spent in education, and high socioeconomic status (Wong et al. 2002; Rose et al. 2008; He et al. 2015; Mountjoy et al. 2018). Multiple lines of evidence for gene-gene or gene-environment interactions in myopia pathogenesis have been identified in the previous studies (Verhoeven et al. 2013a; Fan et al. 2014; Tkatchenko et al. 2015; Pozarickij et al. 2019).

To date, 25 myopia loci have been identified via linkage analyses, and multiple

candidate genes inside the linkage interval have been analyzed (Table 1.2). In recent years, a large number of genome-wide association studies (GWASs) and a series of follow-up association studies have been conducted in different ethnic population (Cai et al. 2019). Single nucleotide polymorphisms are the dominant genetic variants in these studies. The association of numerous variants with myopia-related phenotypic traits, such as refractive error, axial length, and macular thickness, have been repeatedly found (Kiefer et al. 2013; Verhoeven et al. 2013a; Shah et al. 2018).

In this study, an assumption of the high myopia case-control study was that the associated genetic variants have effects on the trait (refractive error) of the same magnitude, irrespective of other variants and other causal factors. This assumption implied, for example, that a specific genetic variant associated with a -2.00 D shift in refractive error in a person who would otherwise have been emmetropic, would also shift refractive error in the direction of myopia by -2.00 D in a person who would otherwise be a +3.00 D hyperope. Therefore, by selecting participants with hyperopia as controls, rather than participants with emmetropia, the difference in phenotype between the case and control group was more extreme. This provided greater statistical power than an analysis of high myopia cases versus emmetropic controls, if the assumption was correct. Since the nature of the trait was binary, logistic regression was chosen as the statistical model for the association study.

STR expansions are difficult to detect and may explain part of the “missing heritability” of high myopia. STRs may vary in length at the level of the individual patient and be subject to distinct selective pressures (Wren et al. 2000; Hannan 2018). The high variability of tandem repeats has been attributed to events including strand-slippage during replication (Pumpernik et al. 2008), retrotransposition (Sulovari et al. 2019), unequal crossing over in meiosis (Gwiazda et al. 2000), and issues in DNA repair (Usdin et al. 2015). STR length variation is identifiable but not commonly analyzed in short-read whole-genome sequencing or whole-exome

sequencing data. In this study, I used the same bioinformatic methods described in Chapter 3 to extract the STR length information from WES data for UK Biobank participants. The STR genotype information was analysed to investigate the association of STR length with high myopia.

4.2. Methods

4.2.1. Selection of Participants

The GWAS and subsequent analyses were restricted to unrelated UK Biobank participants of European ancestry who were part of the October 2020 (WES 200k) data release. The genetic ancestry principal components provided by Bycroft et al. (2018) were used to define a cluster of individuals with European ancestry. Participants whose genetic ancestry PCs did not cluster with Europeans were excluded. Only individuals who did not withdraw their consent were studied. Moreover, participants with a mismatch between their self-reported and genetically-inferred sex, or whose imputed genotype data showed high heterogeneity (heterozygosity >4 standard deviations from the mean level), were also excluded. Genetic markers were genotyped with either the UK BiLEVE and UK Biobank Axiom Array. Only variants that present on both arrays were retained (Bycroft et al. 2018). The markers that failed in more than one batch, had a greater than 5% overall missing rate, or had a MAF < 0.0001 were excluded (Bycroft et al. 2018). Samples that were identified as outliers for heterozygosity or missing rate were also excluded (Bycroft et al. 2018). Genotypes of the individuals were imputed into the dataset using the IMPUTE4 software, with a combined Haplotype Reference Consortium (HRC) and UK10K haplotype reference panel (Bycroft et al. 2018). The reference panel included approximately 96 million variants (in GRCh37 coordinates). The imputed genotypes were aligned to the positive strand of the reference strand, and

imputation was carried out in chunks of approximately 50,000 imputed markers, with a 250 kb buffer region and on 5,000 samples at a time (Bycroft et al. 2018). After applying these filters, there were n=163,340 individuals remaining. Next, participants who self-reported a history of eye trauma resulting in loss of vision, cataract surgery, laser eye surgery or corneal graft surgery were excluded, as were individuals whose hospital records (ICD10 codes) indicated a history of cataract surgery, eye surgery, retinal surgery, or retinal detachment surgery. Participants without a valid mean spherical equivalent autorefraction measurement or who were recruited from a UK Biobank Assessment Center at which the number of the valid samples was less than 50 were also excluded. After applying these criteria, there were 54,204 individuals remaining.

Then, individuals who had an average mean spherical equivalent refractive error for their two eyes of $\leq$ -6.00 diopters or $\geq$ +2.00 diopters were classified as high myopia "cases" and moderate hyperopia "controls", respectively. This resulted in a sample of n=2,005 high myopia "cases" and n=6,928 high hyperopia "controls". From amongst these remaining participants, the maximum sized set of unrelated participants was selected using the method of Bycroft et al. (2018). This led to a final sample comprising of 2,002 cases and 6,806 controls (Fig. 4.1).

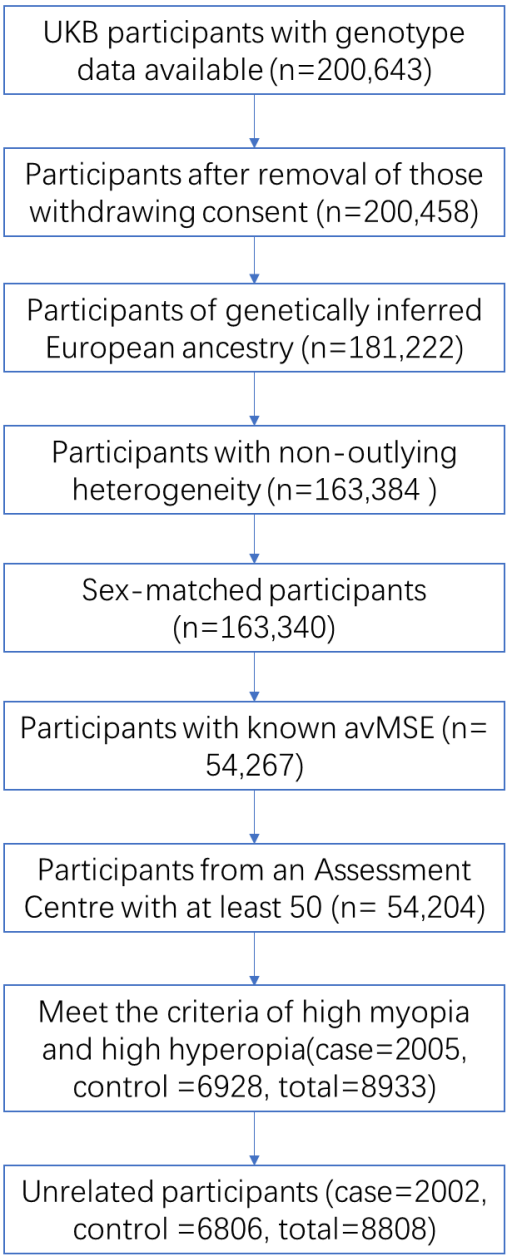


Fig. 4.1 Flow diagram illustrating the selection of UK Biobank participants for the GWAS sample.

4.2.2. Selection of STRs that could be genotyped reliably

The Genome Reference Consortium Human Build 38 (GRCh38) reference panel includes 1,638,945 STRs across all 22 autosomal chromosomes and the X chromosome. As described in section 3.2.2., a two-step process was developed to identify and genotype STRs located within the sequenced regions in the UK Biobank WES data. In the first step, the HipSTR program was used to evaluate all 1,638,945 candidate STRs in a subsample of n=200 randomly selected participants from the full sample of 8,808 individuals. To facilitate parallel computation, the 1,638,945 STRs were split into 16,240 groups, each containing 100 STRs (except for 23 groups containing the last set of STRs on each chromosome). The first step identified 22,711 STRs that could be successfully genotyped using WES data from UK Biobank participants. In the second step, these 22,711 STRs were genotyped in the full sample of 8,808 participants. Since a large number of STRs were genotyped in the second step, this step was parallelized as well. Candidate STRs were split into 255 groups, each containing 100 STRs except for 23 groups containing the last set of STRs on each chromosome.

The HipSTR program (Willems et al., 2017) was used for STR genotyping. Genetic data for the 8,808 UK Biobank participants was input to HipSTR in CRAM format (--bam-files command in HipSTR), alongside a reference assembly GRCh38 (--fasta command in HipSTR). In order to reduce the chance of any genotyping errors, a sufficient coverage of each STR was set. In the first step of the 2-step strategy, STRs genotyped by HipSTR in the sample of 200 participants were required to have a minimum of 1,000 reads (equivalent to 5 reads per participant) and a maximum of 90,000 reads (equivalent to 450 reads per participant). Therefore, in the second step, since the number of samples was larger, the variation of STR coverage was larger as well. In order to successfully obtain genotype information in the larger sample, the above thresholds were loosened. Specifically, for step 2, genotyped STRs were required to

have a minimum of 4,200 reads (equivalent to 0.48 reads per participant) and a maximum of 4,200,000 (equivalent to 480 reads per participant).

4.2.3. Genome-wide Association Study for High Myopia

A genetic association study for high myopia was carried out in the sample of 2,002 cases and 6,806 controls. A total of 15,568 STRs (from the total of 22,403 STRs on the 22 autosomal chromosomes and X chromosome that could be successfully genotyped) were tested for association with high myopia by using logistic regression. The remainder of the STRs exhibited no variation in allele length in this cohort, i.e. they were monomorphic, or they were located outside of exons and therefore could not be genotyped. The average length of the two alleles of each STR genotype was used as the STR predictor variable. Sex, age, age-squared, and the first 10 ancestry principal components were included as covariates. Logistic regression models were fitted using the glm function in R. This approach of using the average length of an STR in a regression analysis was described in section 3.2.3.

In order to increase the power of the genome-wide association study, I estimated the error rates of the STRs by examining the genotype calls of STRs located on chromosome X in males. Based on logic, if the STRs are genotyped accurately, all the alleles on chromosome X should be called as homozygous in males. Hence, the heterozygous allele calls can be directly attributed to the error generated during the genotyping procedure. (However, as discussed in section 4.4, this approach has limitations as a method for assessing the genotyping error rate).

Weighted Bonferroni correction was applied to the p-values from the logistic regression. As with a conventional Bonferroni correction, a weighted Bonferroni correction will control the experiment-wise Type I error rate (α) due to multiple

testing. However, the weighted method offers greater statistical power compared to standard Bonferroni correction when tests can be grouped *a priori* as having a relatively higher or lower chance of success (Rubin et al. 2006; Wasserman and Roeder 2006). Because of the variable genotyping error rates identified for STRs with a motif length of 1 compared to STRs with longer motif lengths (see Results), STRs were separated into two groups: a single-basepair motif group and a multi-basepair motif group. The risk of a Type I error was expected to be higher among the single-basepair motif group STRs due to their higher genotyping error rate compared to the multi-basepair motif group. Thus, the relative weights assigned during the multiple-testing correction step were set as 0.75 and 0.25 (a 3:1 weighting) for the multi-base-pair group and the single-base-pair group, respectively. (In other words, instead of following the convention of setting $\alpha = 0.05\ /n$ when testing all n STRs, alpha was set as $\alpha = 0.05\ \times\ w_j/n_j$, where w_j is the weighting factor for group *j* and n_j is the number of STRs in group *j*, and $\sum w_j = 1$). For a conventional Bonferroni correction, the genome-wide significance threshold for testing 15,568 STRs would be p < 3.21e-06 (=0.05/15,568). Here, the weighted Bonferroni correction approach set the genome-wide significance threshold as p < 1.53e-06 for the single-basepair motif group and p < 5.06e-06 for the multi-basepair motif group. The weighted Bonferroni correction is a robust procedure because the informative specification of the weights can increase power substantially, whereas the uninformative weighting results into little power loss for sparse weights (Roeder et al. 2006; Roeder and Wasserman 2009). Note that the choice of a 3:1 weighting was chosen arbitrarily.

4.2.4. Post-GWAS Analyses

Two STRs associated with high myopia were identified in the current work: Human_STR_827099 on chromosome 2 and Human_STR_424816 on chromosome 14. Conditional association tests were carried out to determine if these two STRs

were independently associated with high myopia after accounting for the association of nearby SNPs. For this test, a conditional logistic regression analysis was conducted. Imputed SNP genotypes for the 8,808 cases and controls were converted from BGEN format to “raw” (R-readable) dosage format using Plink2.0 (Chang et al. 2015) for 20,854 SNPs around Human_STR_827099 (chr2: 232,943,615 – 233,543,615; GRCh37 build coordinates) and 12,349 SNPs around Human_STR_424816 (chr14: 60,574,249 – 61,174,249; GRCh37 build coordinates). The nearby SNPs were included with vs. without the lead STR in the conditional association test, with case-control status as the outcome. Age, age-squared, sex, and the first 10 principal components were included as additional covariates.

4.3. Results

4.3.1. Validation of High Myopia and Hyperopia in UK Biobank Cohort

Among the 8.808 case and control participants, the cases with high myopia had a 5.1-fold lower prevalence of self-reported unilateral amblyopia, a 1.6-fold higher prevalence of 1.00 D or more anisometropia, and an 8.2-fold lower prevalence of self-reported strabismus compared to controls (Table 4.1; Fig. 4.2). Moreover, individuals in the case group had a higher prevalence of early age-of-onset of wearing glasses (age started wearing glasses ≤7 years): 25.1% vs. 18.5% in cases vs. controls). Furthermore, a 1.5-fold lower proportion of cases had a large difference of visual acuity between both eyes (visual acuity difference ⩾2/0 logMAR in both eyes): 18.6% vs. 27.2%. The sex and age of cases and controls also differed subtly between cases and controls, while their Townsend Deprivation Index, a measurement of socioeconomic status, did not show a significant difference (-1.95 vs. -2.14; p= 0.41). The self-reported strabismus showed a significant difference between case and

control group（1.0% vs. 8.2%; p<2.2E-16）, which was identical to the strabismus study in section 3. This information was provided in order to provide consistency between the chapter on Strabismus and the chapter on Myopia case-control status.

Table 4.1 Demographic and ocular characteristics of the UK Biobank strabismus case-control sample

Variable	Statistic	Total (n=8,808)	Cases (n=2,002)	Controls (n=6,806)	*p*-value
Female	N (%)	4,945 (56.1%)	1,182 (59.0%)	3,763 (55.3%)	3.2E-03
Self-reported unilateral amblyopia	N (%)	1,008 (11.7%)	57 (2.8%)	971 (14.3%)	3.2E-44
Both eyes VA ≤0.0 logMAR	N (%)	3,299 (37.9%)	751 (38.0%)	2,548 (37.8%)	0.93
VA difference ≥0.2 logMAR	N (%)	2,178 (25.2%)	363 (18.6%)	1,815 (27.2%)	2.7E-14
Better VA≤0.0; VA difference≥ 0.2	N (%)	1,481 (17.2%)	227 (11.6%)	1,254 (18.8%)	2.8E-13
Anisometropia≥ 1.00 D	N (%)	2,725 (31.0%)	879 (43.9%)	1,846 (27.1%)	4.3E-46
Anisometropia≥ 2.00 D	N (%)	1,176 (13.4%)	359 (18.0%)	817 (12.0%)	9.3E-12
Self-reported strabismus	N (%)	579 (6.6%)	21 (1.0%)	558 (8.2%)	<2.2E-16
Age (years)	Median (IQR)	62.1 (55.8 to 60.2)	56.2 (49.3 to 55.6)	63.2 (58.4 to 61.6)	<2.2E-16
Refractive error (D) average of 2 eyes	Median (IQR)	+2.60 (2.04 to 3.62)	-7.48 (-8.95 to 6.67)	2.95 (2.39 to 4.01)	<2.2E-16
Anisometropia (D)	Median (IQR)	0.57 (0.24 to 1.22)	0.87 (0.39 to 1.63)	0.51 (0.21 to 1.07)	<2.2E-16
Age started wearing glasses (years)	Median (IQR)	20.00 (9.00 to 43.00)	10.00 (7.00 to 13.00)	35.67 (11.00 to 45.00)	<2.2E-16
Townsend Deprivation Index	Median (IQR)	-2.09 (-3.56 to 0.39)	-1.95 (-3.49 to 0.50)	-2.14 (-3.58 to 0.33)	0.41

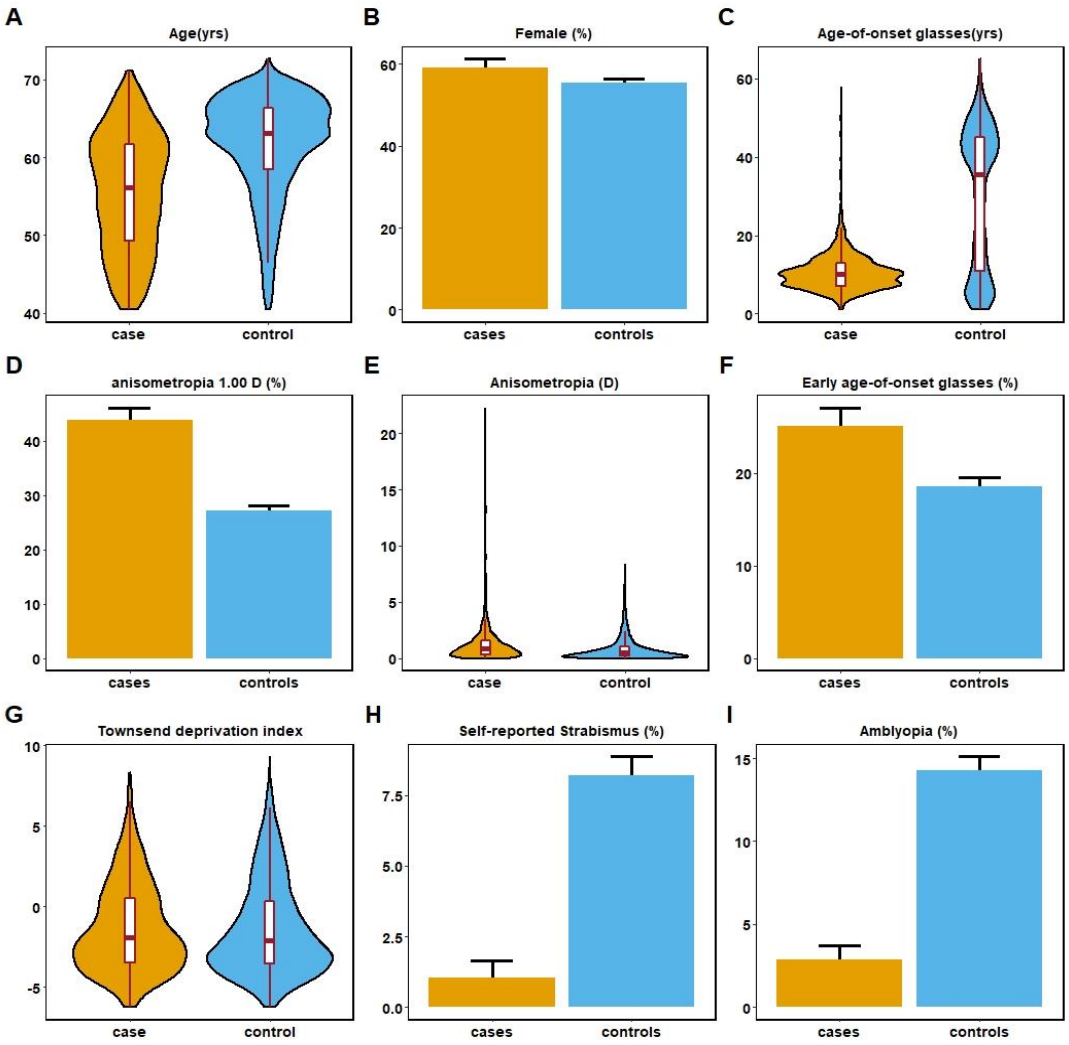


Fig. 4.2 Demography and clinical characteristics of cases and controls (n=2,002 cases and n=6,806 controls). (A) Violin plot for the age of participants. (B) Bar chart for the percentage females. (C) Violin plot for the age-of-onset wearing glasses. (D) Bar chart for the percentage of anisometropia above 1.0 D. (E) Violin plot for anisometropia of participants. (F) Bar chart for the percentage of early age-of-onset of glasses. (G) Violin plot for Townsend Deprivation Index of participants. (H) Bar chart for the percentage of self-reported strabismus. (I) Bar chart for the percentage of amblyopia. Early age-of-onset of glasses was defined as $\leqslant$7 years. Bar chart error bars denote 95% confidence interval. Box plots are superimposed over the violin plots.

4.3.2. Calling Rate Assessments

Only the successfully genotyped STR were included in the call rate assessment. After the two-step HipSTR genotyping process, 22,403 out of the total 1,638,945 STRs from the GRCh38 could be reliably genotyped. To study the association between a phenotype and STR genotypes, the STRs must be polymorphic. Surveying the number of alleles of each STR revealed that 19,850 out of the 22,403 STRs were polymorphic. The 2,553 monomorphic STRs were excluded from further analysis.

Compared to the per-sample call rate, the per-STR call rate displayed relatively greater variation (Fig. 4.3A). Among the full set of 19,850 polymorphic genetic variants, 15,568 STRs had a per-STR call rate above 95%, while 2,229 STRs had a per-STR call rate lower than 90%. The 15,568 STRs with per-STR call rate above 95% were taken forward for use in the GWAS analysis.

After the exclusion of STRs with a per-STR call rate <95%, the per-sample call rate was examined in the set of 8,808 participants. The per-sample call rate exhibited lower variation ($F = 2.2E\text{-}3$, $p < 2.2E\text{-}16$; Fig. 4.3B) but a higher average level compared to the per-STR call rate (98.8% vs. 95.0%, $t = 46.3$, $p < 2.2E\text{-}16$). 8,807 samples had a call rate greater than 95%, and all 8,808 samples had a call rate greater than 93%. To keep the maximum power of the further statistical analysis, all 8,808 samples were included in the further study.

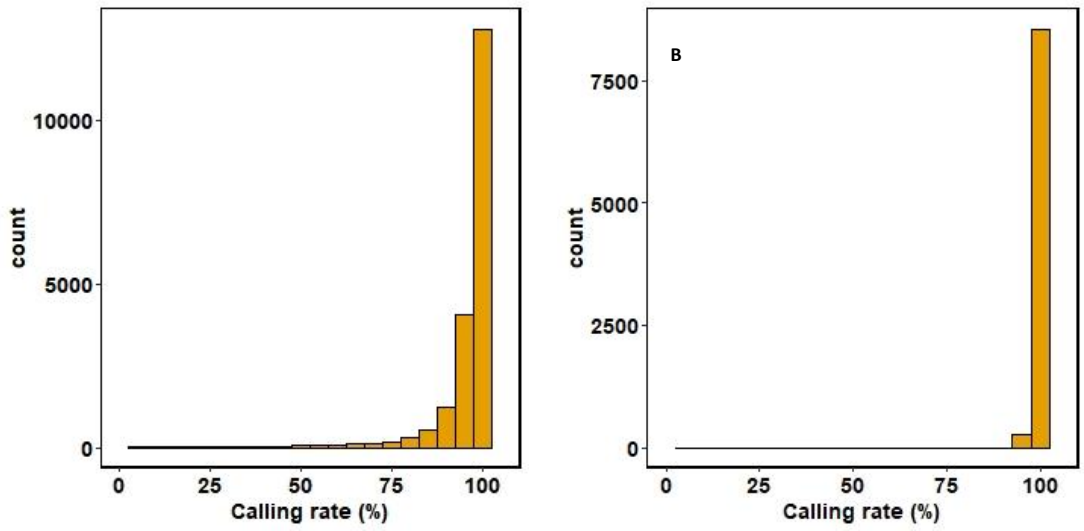


Fig. 4.3 Histogram of per-STR and per-sample call rates. (A) Distribution of per-STR call rate for 19,850 STRs located across 22 autosomes and chromosome X. (B) Distribution of per-sample call rate for 8,808 individuals (after the exclusion of STRs with a per-STR call rate <95%).

Further analysis revealed that the 19,850 STRs were located both inside and outside exon regions: 6,620 STRs were located within exons and 13,230 were situated outside exon regions. Read coverage was also considered, i.e. how many reads were obtained for each sample on average. Unexpectedly, the average per-STR call rate of STRs in exon regions was lower than the call rate of STRs located outside exons: inside vs. outside mean call rate = 0.94 vs. 0.96, respectively, $p < 2.2E-16$; Fig. 4.4 A. However, the mean read count for STRs within exon regions was significantly higher than for those outside exon regions: inside vs. outside mean read count = 56.3 vs. 25.7, $p < 2.2E-16$ (Fig. 4.4 B), suggesting the genotype results of STRs in the exon regions may be more reliable. Shapiro-Wilk's test revealed the distribution of call rates was non-normal ($p < 2.2E-16$); therefore, the correlation between call rates and read counts was estimated with Spearman's rank correlation coefficient. This revealed a negative correlation between ranks of call rates and ranks of read counts ($\rho = -0.10$, $p < 2.2E-16$; Fig. 4.5 A), but the fact that the correlation coefficient was close to zero indicated that the correlated trend was not very strong.

The relationship between the call rate and other STR features was also studied. Since the distribution of call rates was non-normal, the call rate for the groups by the length of the motif was compared using the Kruskal-Wallis rank sum test. A significant difference in the mean call rate was found ($p < 2.2E-16$; Fig. 4.4 C). The 5-bp motif group had a lowest call rate (92.5%), while the 2-bp motif group had the highest call rate (96.5%). The 5-bp motif group had a significantly different mean call rate compared to the other motif groups (Wilcoxon rank sum test $p1,5 = 5.5E-9$, $p2,5 = 7.2E-3$, $p3,6 < 2.2E-16$, $p4,5 = 4.0E-11$, $p5,6 = 3.3E-09$). Meanwhile, the mean read count per sample for groups by the length of motif was also compared using the Kruskal-Wallis rank sum test. A significant difference in the mean call rate was found ($p < 2.2E-16$; Fig. 4.4 D). The 3-bp motif group had the highest read count (70.0 reads), and the 1-bp motif group had the lowest read count (24.6 reads).

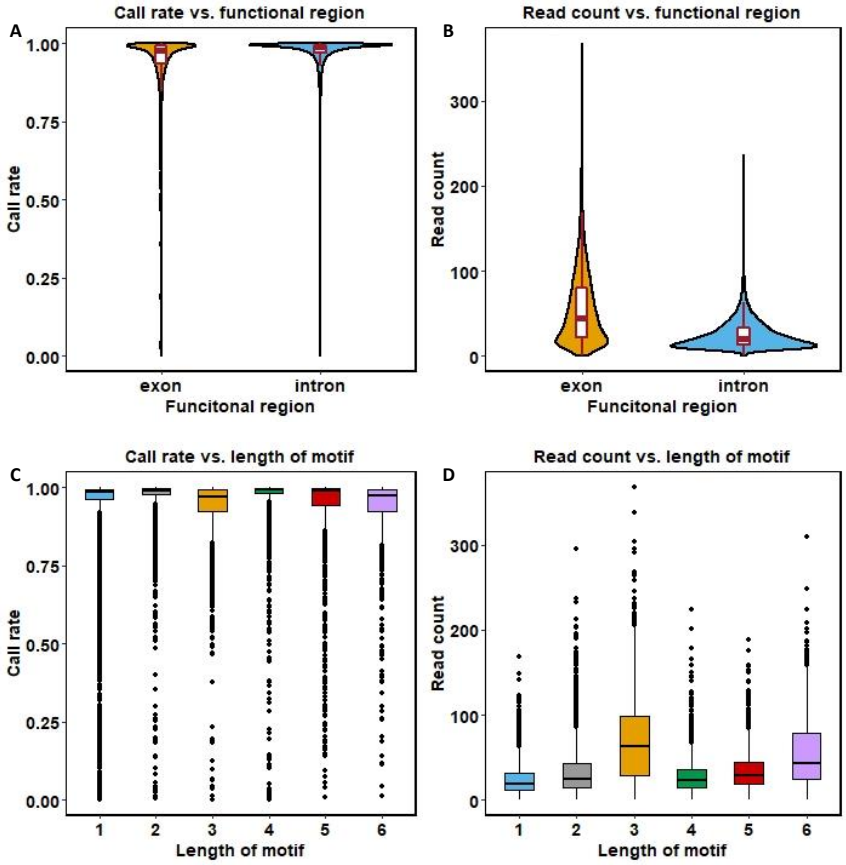


Fig. 4.4 The quality of genotype information for 19,850 STRs. (A) The distribution of call rate versus functional region. (B) The distribution of read count versus functional region. (C) Graph of call rate versus motif length. Box plots are superimposed over the violin plots. (D) Graph of read count versus motif length.

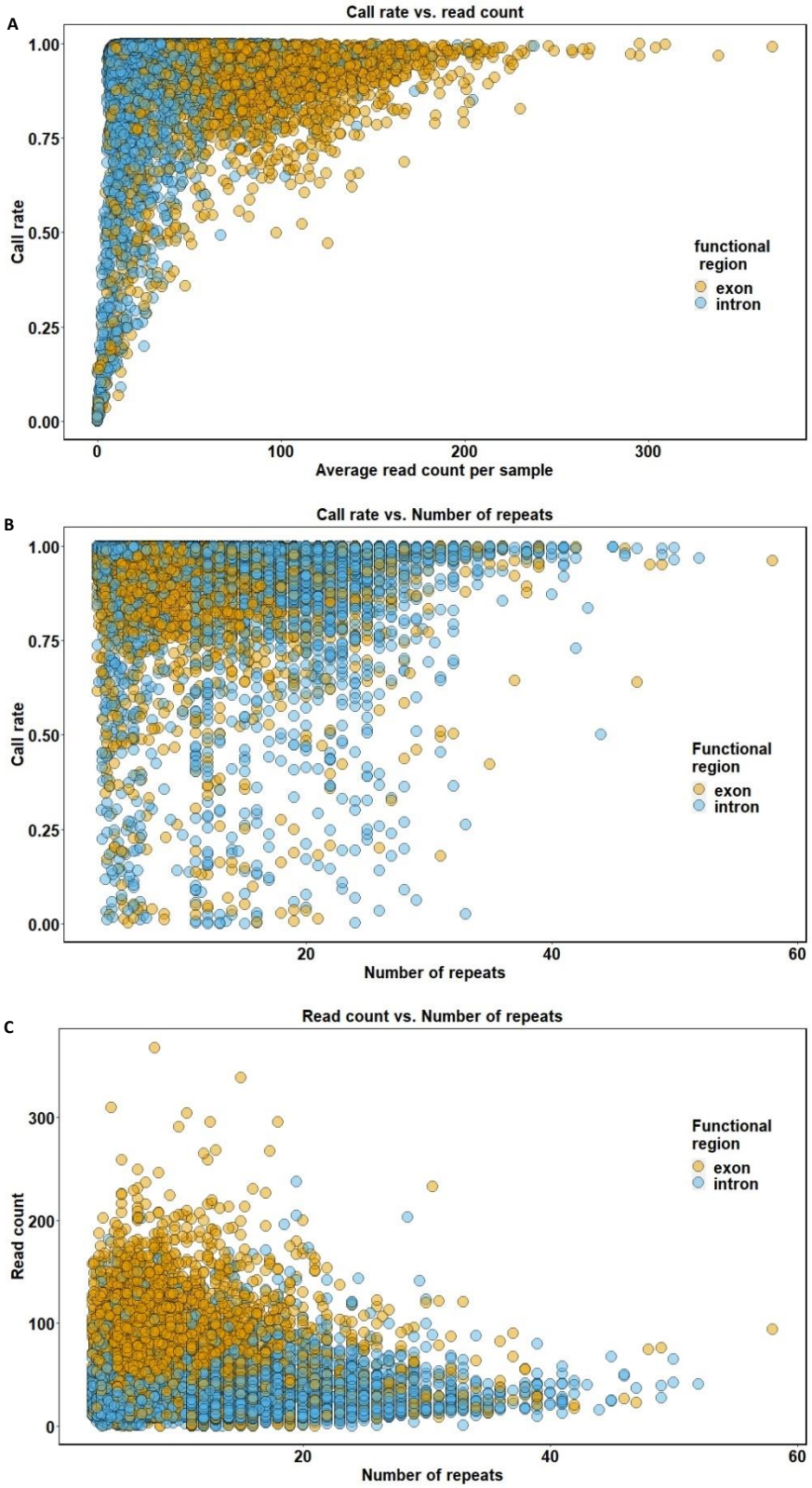


Fig. 4.5 The quality of genotype information for 19,850 STRs (continued). (A) The relationship between call rate versus read count. (B) The relationship between call rate versus referenced number of repeats. (C) The relationship between read count

versus referenced number of repeats. Different colors represent different functional regions.

A negative correlation was found between the call rate versus the observed number of repeats of STRs (ρ = -0.17, p < 2.2E-16; Fig. 4.5 B). A further negative correlation was found between the read count per sample versus the number of "referenced repeats", which corresponds to the number of repeats in the GRCh38 reference genome (ρ = -0.15 p < 2.2E-16; Fig. 4.5 C).

4.3.3. Estimation of the Genotyping Error Rate

The genotyping error rate (the proportion of observed genotypes versus true genotypes) of genetic markers can be adversely affected by various causes (Bonin et al. 2004). Here, the genotyping error rate of the STRs was estimated by quantifying the rate of heterozygous genotype calls for alleles on chromosome X among male individuals. Because alleles on chromosome X should always be called as homozygous in males, every heterozygous call indicates a genotyping error.

The average heterozygous genotype call rate was calculated for alleles grouped by motif length (Table 4.2; Fig. 4.6). The 1-bp motif group had the highest error rate, which was significantly higher than the other groups (*t*-test p1,2 < 2.2E-16, p1,3 < 2.2E-16, p1,4 < 2.2E-16, p1,5 < 2.2 E-16, p1,6 =1E-14). Meanwhile, the error rate of the 2-bp motif group was significantly higher than the 3-bp and 4-bp motif groups (*t*-test p2,3 < 7.3E-4, p2,4 < 9.1E-3).

Table 4.2 Heterogeneous rate for STRs on chromosome X within male individuals

Length of motif	Number	Heterozygous call rate (%)	95% C.I.
1	234	26.5	24.6,28.5
2	92	7.8	5.9,9.7
3	92	2.1	1.5,2.6
4	48	2.5	1.6,3.3
5	23	4.6	1.2,8.0
6	20	5.2	0.2,10.2

Abbreviations: Number = number of STRs; 95% C.I. = 95% confidence interval.

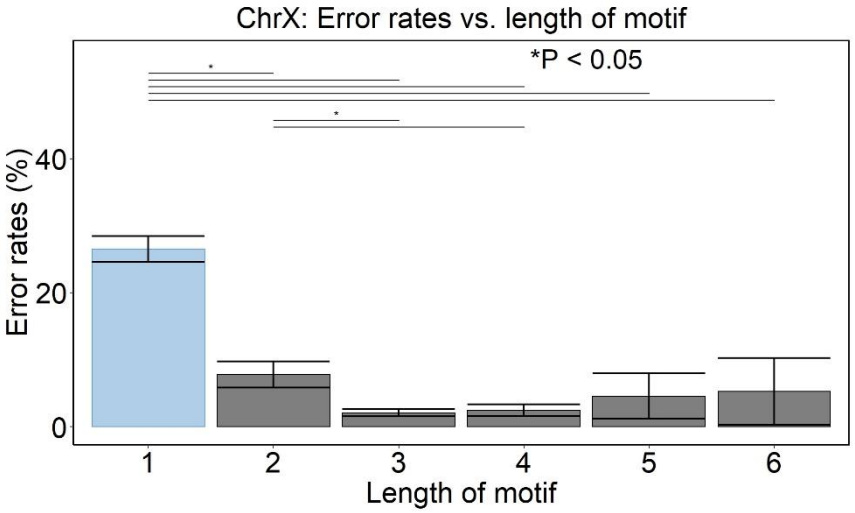


Fig. 4.6 Bar plots for the error rates versus length of motif on chromosome X. The error bars of the bar plots represent the 95% confidence interval.

4.3.4. STR-based GWAS for high myopia case-control status

A logistic regression analysis testing for association of the high myopia case-control status and the average length of the 2 STR alleles carried by each participant was carried out. The GWAS included the 15,568 polymorphic STRs with a per-STR call rate >95% that were distributed across all 22 autosomal chromosomes and the X chromosome. A weighted Bonferroni correction method was applied; the overall alpha for Bonferroni correction is $\alpha = 0.05\ /n$ = 3.21E−06, where n is the number of the STRs (n =15,568). The numbers of STRs with single-bp and multi-bp motifs was 8,156 and 7,412, respectively. After applying a weighting scheme of 3:1 to the multi-bp motif group and the 1-bp motif group, the p-value thresholds for declaring statistical significance were set as: α =1.53E-06 for the 1-bp motif STRs, and $\alpha =$ 5.05E-06 for the multi-bp motif STRs.

The GWAS identified two STRs that were significantly associated with high myopia case-control status: STRs Human_STR_827099, OR = 0.67 (95% CI = 0.57 to 0.79, p = 6.5E-07) and Human_STR_424816, OR = 0.69 (95% CI = 0.59 to 0.80, p = 1.5E-06), located on chromosomes 2 and 14, respectively (Table 4.3; Fig. 4.7). Focusing on the covariates, a negative association with high myopia was identified for age: OR = 0.36 (95% CI = 0.33 to 0.39, p < 2.2E-16) and age-squared: OR = 0.19 (95% CI = 0.74 to 0.84, p = 1.0E-11).

In order to check for systematic bias due to population stratification, the genomic inflation factor (λ_{GC}) was calculated. There was no evidence of population stratification ($\lambda_{GC} = 0.87$). A Q-Q plot is presented in Fig. 4.8.

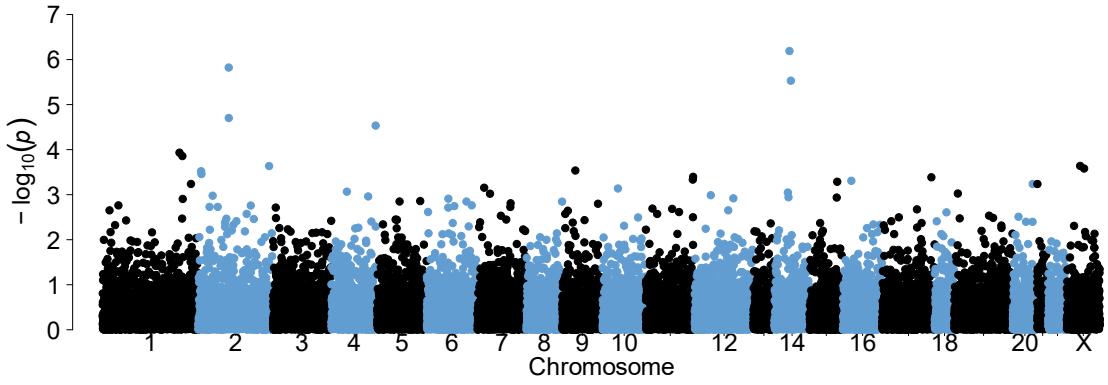


Fig. 4.7 Manhattan plot for STR-based GWAS for high myopia case-control status.

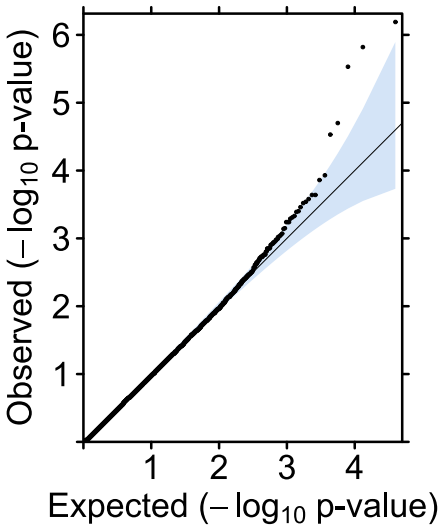


Fig. 4.8 Quantile-quantile (Q-Q) plot for p-value of logistic regression analysis. The x-coordinate denotes the theoretical quantile for a uniform distribution, and y-coordinate denotes the observational sample quantile distribution. The blue shaded region denotes the confidence interval of the theoretical distribution.

Table 4.3 Lead STR variants associated with high myopia case-control status

STR	CHR	Start	End	Motif	Repeats	p-value
Human_STR_827099	2	233,243,615	233,243,630	AC	8.0	1.5E-06
Human_STR_424816	14	60,864,407	60,864,417	A	11.0	6.5E-07

Abbreviations: CHR: chromosome; Start: the starting locations for the STRs in the GRCh37 build; End: the ending locations for the STRs in the GRCh37 build; Repeats: the number of repeats of the motif in the reference exome library.

4.3.5. Regional GWAS analysis on chromosomes 2 and 14

The Human_STR_827099 variant is located in the first intron of the *ALPP* gene, and the repeat has a repetitive 'AC' motif (except the insertion of adenine in the fifth nucleotide position). Hence, the repeat would not be expected to change the amino acid sequence of the gene or block RNA polymerase binding sites. The product of the *ALPP* gene is alkaline phosphatase, which can hydrolyze various phosphate compounds. Alkaline phosphatase is the isozyme that exists in the placental of most mammals; also, it exists at high concentrations in the liver and bones. No publication to date has reported a link between *ALPP* and myopia or refractive error.

The most common minor allele of Human_STR_827099 has one more repeated 'AC' motif than the GRCh38 reference allele. There is no accessible record of the minor allele frequency of this variant in genetic libraries. In the 6806 controls with European ancestry, the frequency of this 1-motif expanded allele is 0.087, which is within the class of 'common variant'.

The Human_STR_424816 variant is located in the tenth intron of the *RBM8B* pseudogene. The GRCh38 reference allele has 11 repeats of an 'A' motif. *RBM8B* codes for the RNA binding motif protein 8B pseudogene. The highly conserved RNA-binding motif protein produced by the *RBM8B* gene was found to interact with *OVCA1*, a candidate tumour suppressor (Salicioni et al. 2000).

The most common minor allele of Human_STR_424816 has one less 'A' repeat than the GRCh38 reference allele. There is no accessible record of the minor allele frequency of this variant in genetic libraries. In the 6806 controls with European ancestry, the frequency of this 1-motif expanded allele is 0.21, which is also within the class of 'common variant'.

Linkage disequilibrium between the two STRs and nearby SNPs known to be associated with refractive error was examined. Guggenheim et al. (2022) identified SNPs in the *PRSS56* gene, predicted to have functional consequences, that were associated with refractive error in GWAS studies. Human_STR_827099 is located approximately 1Mb downstream of the *PRSS56* gene. Similarly, Guggenheim et al. (2022) also reported functional SNPs in the *SIX6* gene associated with refractive error. The Human_STR_424816 variant is located 400kb downstream of *SIX6*. Given the proximity of the physical location of these two STRs to known refractive error candidate genes *PRSS56* and *SIX6*, the extent of linkage equilibrium and further conditional association analysis were performed.

The independent roles of the STRs and the nearby genes onto the refractive error was examined. Guggenheim et al. (2022) reported functional annotation SNPs in the PRSS56 gene, which was identified the link to the refractive error. The location of the Human_STR_827099 variant is within 1Mb downstream of PRSS56 gene. Similarly, functional annotation SNPs in SIX6 gene was identified the association with refractive error. The location of the Human_STR_424816 variant is within 400kb downstream of SIX6 gene. Given the proximity of the physical location on the chromosome 2, the extent of linkage equilibrium and further conditional association analysis were applied.

The independence of the relationship with high myopia case-control status between each lead STR and nearby SNPs was tested by performing a conditional analysis. Regional conditional GWAS analyses were performed for regions spanning ±300 kb from the lead STRs. To provide a high SNP density for the regional GWAS analyses, the SNPs were imported from the UK Biobank imputed SNP dataset, which gave an increased number of SNPs for the regional conditional analysis. The same GWAS parameters for SNPs and covariates were adopted as in the original GWAS except that the genotype of the lead STR was included as an additional covariate. The

Liftover program was used to convert between GRCh37 genomic coordinates (imputed genotypes) and GRCh38 genomic coordinates (WES genotypes).

A conditional analysis was performed for all SNPs located ±300 kb from the lead STR. 20,854 SNPs near the STR Human_STR_827099 and 12,349 SNPs near the STR Human_STR_424816 were included in the conditional analyses. The LD between each SNP and the lead STRs was calculated as the correlation of the alleles. For both loci, the conditional analysis that included the lead STR produced a clear increase of the p-values for SNPs in LD with the lead STR, whereas the other SNPs retained their original p-values for association (Fig. 4.9). The change in p-value for the lead independent SNPs are listed in Table 4.4.

Because of the different approaches of representing the genotype of alleles, the p-values for SNPs and lead STRs are not directly comparable by the current method. Further study will be required to determine if the two lead STRs influence refractive development directly or if they tag the causal variant(s) in these regions.

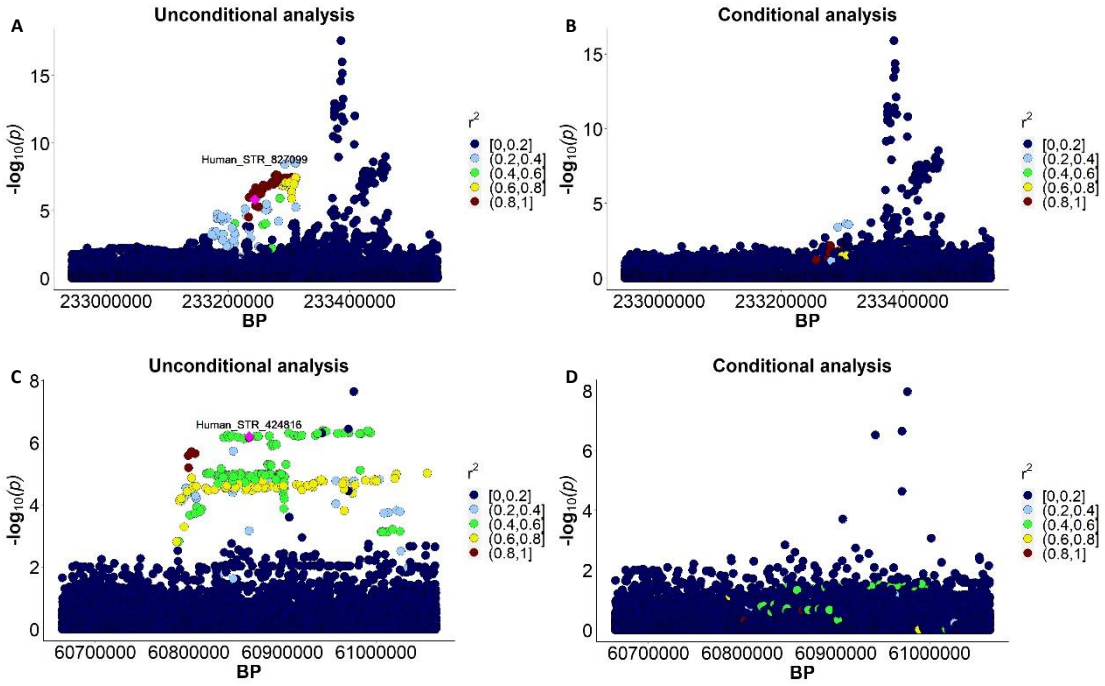


Fig. 4.9 Regional association plot for SNPs in the region of strongest association, before and after conditioning on the lead STR. Panel (A) and (B) before and after conditioning on Human_STR_827099 (chromosome 2). Panel (C) and (D) before and after conditioning on the lead STR Human_STR_424816 (chromosome 14). The lead STR in the region is indicated by a purple diamond. In the conditional analysis, the average allele length of the lead STR was included as a covariate in the logistic regression model, along with the baseline covariates. If the inclusion of the STR genotype in the conditional analysis has minimal impact on the strength of association for the lead SNP (i.e. the p-value for the SNP remains similar in the baseline and conditional analyses), then this implies that the association of the SNP and the trait is independent of the STR genotype. By contrast, if the inclusion of the STR genotype in the conditional analysis has a major impact on the strength of association for the lead SNP (i.e. the p-value for the SNP is shifted towards the null), this implies that the SNP and the STR are surrogates for each other and thus that they are not independently associated with the trait.

Table 4.4 Conditional analysis results for the lead SNPs

Nearest Gene	Variant	rsID	MAF	P_{uncond}	P_{cond}	LD (r^2)
PRSS56	2:232520686:G:A	rs1550094	0.334	2.5E-18	1.2E-16	0.078
PRSS56	2:232378227:T:C	rs3762525	0.10	7.6E-07	0.99	0.98
SIX6	14:60509783:G:A	rs146737847	0.007	2.3E-08	1.1E-08	0.001
SIX6	14:60397688:TA:T	rs111689247	0.10	6.9E-05	0.20	0.82

Abbreviations: Variant = Chr:Pos:Ref:Alt Genomic position in GRCh38 coordinates; MAF=Minor allele frequency; P_{uncond} = p-value of unconditional analysis; P_{cond} = p-value of conditional analysis; LD = linkage disequilibrium.

4.4. Discussion

This exome-wide STR-based association study revealed an association between high myopia case-control status and specific STR polymorphisms located near to the *PRSS56* and *SIX6* genes. In a SNP-based analysis of WES data from UK Biobank participants, Guggenheim et al. (2022) identified two independent signals in both *PRSS56* and *SIX6*, which were inferred to be driven by different casual variants. The conditional analysis in my current study found the same signals as the Guggenheim et al. study, but using STRs as innovative genetic markers. As described above, a change of *p*-value for the SNP-trait association in the baseline vs. conditional analysis suggests there is a dependent relationship (LD) between the lead SNP and the lead STR. However, due to the complexity of performing a statistical test of association for a multi-alleleic STR – compared to the simplicity of testing for association with a biallelic SNP – the *p*-values quantifying the strength of association for STRs and SNPs are not directly comparable. In other words, these *p*-values are an imperfect guide to the relative importance of the strength of the STR-trait and SNP-trait associations. The presence of both SNPs and STRs in high LD nearby the *PRSS56* and *SIX6* genes made it difficult to pinpoint the most likely causal variants in these regions. Further studies will be required to find out if the STRs influence refractive development directly or if they tag the causal variants in these regions.

The difference for the setting of the minimum reads in the two-step HipSTR was based on empirical optimisation. In these optimisation trials, the proportion of STRs that could be genotyped at scale in step 2 was compared to the number that were genotyped in step 1. Variability in the reads-per-sample across the large sample included in step 2 was reasoned to explain the inconsistency of this step. The numbers of STRs after the genotyping process was 1,197 on chromosome 17, which was consistent with the number of variants after the genotyping in the strabismus

study (section 3.3.2). These findings suggest that the more lenient threshold for the minimum reads-per-sample in step 2 did not adversely affect the accuracy of the genotyping process.

In this study, genotyping accuracy was accessed by using the heterozygous call rate for STRs on chromosome X in males. However, this method could underestimate the true error rate of genotyping on the autosomal chromosomes, because it would not capture all sources of genotyping error (e.g. calling heterozygous genotypes is likely to be more challenging than calling homozygous genotypes when the read count is low). Interestingly, the current analysis revealed only a weak relationship between genotyping accuracy and read depth. Typically, for STRs with more reads, one would expect the genotyping accuracy to be higher and vice versa. For the male samples, the number of reads for each STR on chromosome X was approximately 0.5-fold lower than in females (26.4 vs. 46.1, $p < 2.2E-16$; Fig. 4.10 A). Thus, the genotyping accuracy is likely to be higher in females than males for STRs on chromosome X. On chromosome 1, by contrast, no significant difference in read count by sex was identified between male and female samples (41.8 vs. 40.4, $p = 0.21$; Fig. 4.10 B). Therefore, it is likely that the genotyping accuracy of chromosome X for the male samples is lower than the average level for the autosomal chromosomes. By manually checking the sequence reads on chromosome X, I found that many heterozygous genotype calls were derived from one single read error; with a larger read count, such isolated events are more likely to be interpretated as random noise.

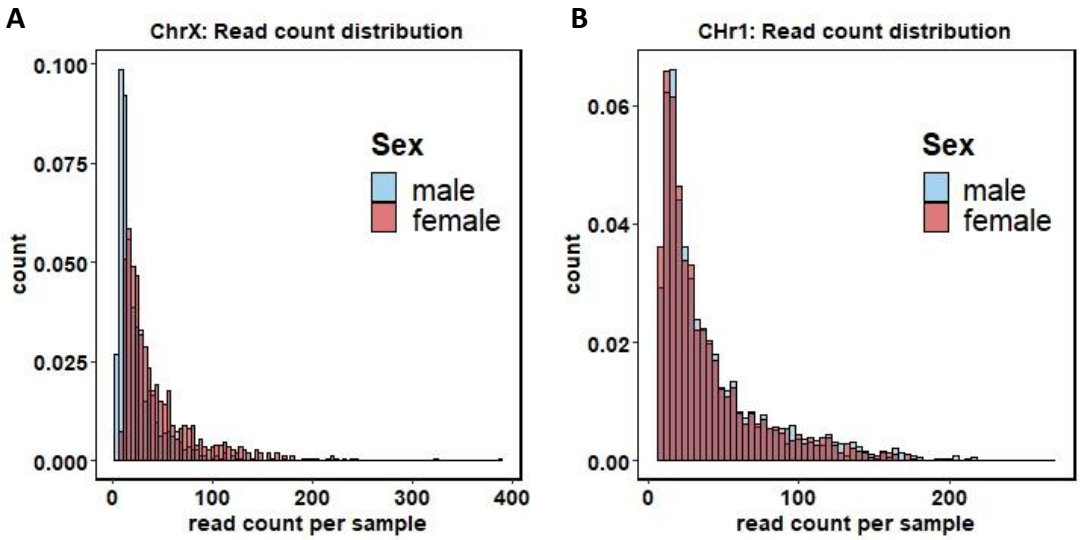


Fig. 4.10 Histograms of read counts of STRs by sex. (A) The density of read counts of STRs on chromosome X. (B) The density of read counts of STRs on chromosome 1.

The choice of moderate-to-high hyperopes as controls in the current study was made based on the assumption that genetic susceptibility would be more easily identified in cases and controls with a more extreme difference in phenotype (compared to studying high myopia cases and emmetropic controls). Hyperopia and myopia represent the opposite arms of the refractive error distribution. Ocular structures such as the thickness of choroidal vascular bed and scleral coat (Sundin et al. 2005), also grow in the opposite way between myopes and hyperopes. Furthermore, previous genetic studies have yielded overlapping regions of association for hyperopia and myopia, providing justification for our method. For example, the 15q14 locus (near *GJD2*) and the 8q12 were found to be associated with both myopia and hyperopia (Kiefer et al. 2013; Verhoeven et al. 2013b; Simpson et al. 2014). The direction of effect is exactly opposite in the myopia and hyperopia studies – suggesting the causal variants operating the mechanism on the whole spectrum of refractive error (Tideman et al. 2021).

In each of the genes *PRSS56* and *SIX6*, I identified two independent signals when studying both STRs and nearby SNPs. One signal was tagged by the lead STR associated with case-control status, while the second signal was tagged by variants not in LD with the lead STR. The same genetic architecture was identified by an independent SNP-based GWAS that included 51,624 unrelated adults of European ancestry from UK Biobank (Guggenheim et al. 2022). The lead variants overlapped between my regional analysis and the previous study (Tables 4.4 and 4.5). The similar findings have therefore confirmed the complex pattern of genetic association of *PRSS56* and *SIX6* and myopia, although as both studies examined data from UK Biobank, the similar results are not independent sources of evidence.

Table 4.5 Fine-mapping of GWAS regions identified using WES data, taken from the article by Guggenheim et al. (2022)

Nearest Gene	Variant	rsID	P	MAF
PRSS56	2:232520686:G:A	rs1550094	1.21E-24	0.334
PRSS56	2:232523470:G:T	rs74703359	1.09E-07	0.002
SIX6	14:60509783:G:A	rs146737847	1.65E-16	0.007
SIX6	14:60509819:C:A	rs33912345	5.29E-11	0.387

Abbreviations: Variant = Chr:Pos:Ref:Alt Genomic position in GRCh38 coordinates; P = p-value of association with refractive error; MAF=Minor allele frequency

There is supporting evidence for a causal role for the secreted trypsin-like serine protease (encoded by the *PRSS56* gene) in refractive error development. The gene was found to be highly expressed in ganglion cells of adult animals (Visscher et al. 2017). Loss of function mutations in the gene are associated with microphthalmia in humans and reduced eye size in knockout mice (Gal et al. 2011; Orr et al. 2011). Specifically, mice carrying a mutation in *PRSS56* exhibit reduced ocular size (Nair et al. 2011) and *PRSS56* mutations lead to nanophthalmos (posterior microphthalmia) and extreme hyperopia characterized by a significant reduction in ocular axial length in humans (Orr et al. 2011). Many GWAS analyses have independently revealed an association between *PRSS56* variants and myopia, which suggests the gene has pleiotropic effects on the development of refractive error (Kiefer et al. 2013; Verhoeven et al. 2013b). My study identified a novel association between an STR genetic marker within the *PRSS56* gene and refractive error. Because the case-control samples were selected among highly myopic and moderately hyperopic participants, the lead STR is likely to have bidirectional effect on both the case and control groups. The negative regression coefficient found in the current study indicates that an increased number of Human_STR_827099 repeats is associated with susceptibility to hyperopia, whereas a reduced number of repeats is associated with susceptibility to myopia.

SIX6 is also involved in ocular morphogenesis (Alfano et al. 2005). The SIX Homeobox 6 (*SIX6*) gene is part of a group of evolutionarily conserved genes, which are known as eye transcription factors (Ledford et al. 2017). Linkage studies suggest the *SIX6* plays a part in the growth of retinal cells (Abu-Amero et al. 2015). As with the *PRSS56* STR, based on the negative regression coefficient of the *SIX6*-associated Human_STR_424816, additional repeats were associated with increased susceptibility to hyperopia, whereas fewer repeats were associated with susceptibility to myopia.

Past GWAS studies have found the majority of the significantly associated variants to be located in non-coding regions of the genome, which makes the identification of disease-susceptibility genes less straightforward. In contrast, the current study focused on whole exome sequencing data, which made it feasible to implicate specific genes as putative myopia susceptibility genes. The replicated discovery of *PRSS56* and *SIX6* by using innovative genetic markers provided additional evidence that these genes can make important contributions to the development of high myopia. By performing conditional analyses, in which the effects of the lead STRs in each region were accounted for, the contribution to the GWAS signal from nearby SNP variants was able to be evaluated (Orozco et al. 2010). One more insight for the selection of STRs as genetic variants was that the polymorphisms of the variants provide novel way to explain the bidirectional effects from the causal genes. As the increase and decrease of the expansion of alleles were both taken into account, the significantly associated STRs indicate the effect spanning the whole spectrum of the trait. Different mutations of the same STR may lead to opposite traits, such as myopia and hyperopia. Therefore, upon the hypothesis of the bidirectional effect of a causal STR, the case/control groups for GWAS can be locked with the opposite phenotypic traits, instead of a typical pathologic/healthy paired group.

This is the first large-scale study to screen for STRs variants associated with refractive error. Notably, the regions identified as being associated with refractive error were already reported in prior GWAS and family-based sequencing studies (Jiang et al. 2015; Sun et al. 2015; Jin et al. 2017). This shows the limited effect any particular genetic region has on high myopia in a specific population. The pathway(s) through which the *PRSS56* and *SIX6* variant impacts refractive error may hold the potential of being a therapeutic target for slowing the progression of myopia.

Chapter 5. General Discussion and Future Work

5.1. Genetic Predisposition to Strabismus

Novel genetic markers (STRs) in the exon regions of chromosome 17 were studied to test for an association with self-reported strabismus. However, no chromosome-wide significant association was identified. Strengths of this study included the careful matching of the case and control individuals during selection of the target population, which reduced the risk of classification bias. Per-sample and per-STR genotyping call rates were monitored in order to reduce the false-positive and false-negative rate. Age and sex were included in the statistical model to eliminate the risk of a spurious association resulting from demographic differences between the case and control groups. Refractive error was included as an additional covariate to eliminate the cofounding bias from the comorbidity of strabismus and hyperopia. Bonferroni correction was performed to account for multiple-comparisons and thus control the experiment-wise Type I error. Finally, a conditional analysis was performed to investigate the independence of the association signals of the lead SNP and lead STR.

In the past two to three decades, genetic studies of strabismus have not gained as much attention as those for other ocular traits such as refractive error and glaucoma, probably due to low prevalence of strabismus and the challenge for diagnosis and quantitative measurement. Population-based genetic association studies for strabismus have only been reported in recent years (Shaaban et al. 2018; Plotnikov et al. 2019). Moreover, family-based and twin studies have provided only limited evidence to support Mendelian Inheritance of strabismus. Beyond that, the genetic contribution to strabismus remains largely unexplained.

My association study of self-reported strabismus in participants of British European

ancestry using STR markers provided an opportunity to better define its genetic architecture and understand its pathological pathways. However, the lack of any significantly associated signal meant that no new insight could be gained. The most likely reason for the lack of a significant association was insufficient statistical power. Plausible approaches in the future to increase the statistical power include: (1) increase the sample size, (2) improve the accuracy of phenotype ascertainment or measurement to reduce noise in classifying cases and controls, and (3) apply more sophisticated statistical models.

Synthetic signal data has been used to improve STR detection and reduce genotyping bias, through recent improvements in base-calling accuracy. De Roeck et al. (2019) proposed a tandem repeat characterization tool called NanoSatellite to analyse data generated by the high-throughput PromethION sequencer. Repeat length accuracy was quantified for both alleles of the *ABCA7* STR (De Roeck et al. 2019), which is a recently discovered STR associated with the risk of Alzheimer's disease (De Roeck et al. 2018). The *ABCA7* STR (chr19:1049437-1050028, hg19) is a 25-bp repeat that has a high GC content, with frequent nucleotide substitutions and insertions: the total repeat size can reach more than 10,000 bp (De Roeck et al. 2019). Meanwhile, Giebelmann et al. (2019) used a hidden Markov model to identify STR regions and upstream/downstream flanking sequences with the help of signal alignment to flanking regions. Their tool, STRique, has been evaluated on GGGGCC repeats such as FTD/ALS synthetic sequences (Giesselmann et al. 2019). Fang et al. (2022) recently published an article reporting the tool, DeepRepeat, which instead of using base-called reads, detects STRs from nanopore sequencing data through direct analysis of electric signals. DeepRepeat allows the analysis of STR within or close to very low-complexity genomic regions, such as telomeric regions (Fang et al. 2022).

An alternative method to control the type 1 error rate is to set a specific false

discovery rate (FDR) for the association results. A disadvantage of the Bonferroni method is that it is a relatively conservative method, which may limit the power of the discovery of true positive results. The FDR method is less conservative. The aim of the FDR approach is to achieve the smallest possible fraction of false signals among all those that are declared to be true. The total number of rejections of the null hypothesis includes both the number of false positive (FP) and true positive (TP). The formula for FDR is FP/(FP+TP). In the controlling procedure, methods for rejecting the null hypothesis were established, such as the Benjamini–Hochberg procedure and Benjamini–Yekutieli procedure, to control the FDR at level alpha. There are precedents for using the FDR approach in GWAS research projects (Nelson et al. 2017; He et al. 2021).

5.2. Genetic Predisposition to High Myopia

The findings from the high myopia case-control study shed the light on the application of using STR markers to refine GWAS signals. By performing an exome-wide association study, I identified two STRs associated with susceptibility to high myopia. The discovery of STR-based associations implicating the genes *PRSS56* and *SIX6* confirmed the associations with SNPs found in a closely-related previous study (Guggenheim et al. 2022).

The current work has paved the way for further STR-based association studies for high myopia. The selection of the case and control group maximized the level of the difference in refractive error, which was expected to have enlarged the effect sizes of the potential causal variants. The sample size was set to be the largest possible, given the current available data from UK Biobank. Additional strengths of the study

included an examination of factors affecting the per-marker and per-sample genotyping call rate, which reduced the risk of spurious association signals. The weighted Bonferroni correction provided higher statistical power than would have been the case if the differential genotyping error rate of single-bp vs. multi-bp STRs had not been taken into account, while still controlling the overall Type I error rate. Finally, the conditional analysis examined the independence of association signals for STRs and nearby SNPs.

An important role for STRs in human disease was established decades ago, with the discovery of pathogenic repeat expansions in Fragile X Syndrome (Kremer et al. 1991) and spinal and bulbar muscular atrophy (La Spada et al. 1991). Despite the clear implication of STRs in disease, they only rarely featured in medical sequencing studies. This is probably because, while next-generation sequencing has the potential to profile the more than one million STRs known to exist, calling STR genotypes from WGS and WES datasets has proven to be challenging (Li 2014b). Indeed, even known pathogenic STR mutations can be missing in most sequencing pipelines (Keogh and Chinnery 2013).

Although STR-based GWAS analyses offer future promise for studying vision-threatening eye diseases, the challenge remains to obtain reliable STR genotypes in samples of many thousands of participants. The HipSTR software was specially developed to deal with genotyping errors and obtain STR genotypes from WGS or WES datasets. Yet, long or complex repeats still fail to be detected or called correctly by HipSTR, which is probably due to the limited depth of sequencing reads available for genotype calling. Also, HipSTR applies a computationally intensive method to optimize the accuracy of genotyping. Therefore, implementing the method can be time-consuming. For example, genotyping whole-genome sequencing data for a sample of 100,000 individuals, which would involve genotyping approximately 50-

times more STRs than the current WES dataset, would take about 3,000 days (using HipSTR without parallelization). To improve future STR-based association studies, more computationally efficient methods need to be developed, such as the approach recently reported by Fearnley et al. (2022). This study proposed a novel STR detection method in which de novo searches for repeat expansions are performed, without the requirement for alignment or specification of the location of putative repeats (Fearnley et al. 2022). Furthermore, repeats with longer motif size (>6 bp) are currently ignored by the NGS algorithms (Gelfand et al. 2014). The limited number of STRs in the human genome means the STR-based GWAS method will inevitably suffer from limited resolution, compared to SNP-based studies. Finally, in the current work, I used the average length of the two STR alleles carried by an individual to represent the STR genotype. Given the complex allele spectra of most STRs, more elegant approaches may offer benefits in the future.

5.3. Future Work

The two GWAS analyses performed using STR markers have validated the hypothesis that STRs in exons are able to detect regions of genetic association with ophthalmic traits. However, given the limitations of my study, such as the limited sample size and the restriction to analyzing STRs located within or adjacent to exons, I would recommend in future the analysis of whole-genome sequencing data in order to broaden the number of markers that can be included. In general, past SNP-based GWAS experiments have shown that signals are more often located in non-coding regions of the genome (Visscher et al. 2017) rather than within exons. Therefore, studying STRs in non-coding regions will increase the chance of identifying novel loci. I would also recommend testing both common STR variants and low-frequency STR

variants. Although, low-frequency STRs are more challenging to genotype, rare variants offer an advantage in more directly implicating disease genes (Guggenheim et al. 2022). Finally, performing STR-based GWAS analyses in non-European populations should be a priority. To date, the great majority of STR-based GWAS have been reported for European samples, which is far from ideal in view of the geographic differences in prevalence for many eye diseases.

Code Availability

shell scripts programmes: https://github.com/J-one-two/GWAS4HM.git

References

Abrahamsson, M. and Sjöstrand, J. 2003. Astigmatic axis and amblyopia in childhood. *Acta Ophthalmologica Scandinavica* 81(1), pp. 33-37. doi: 10.1034/j.1600-0420.2003.00022.x

Abu-Amero, K., Kondkar, A. A. and Chalam, K. V. 2015. An Updated Review on the Genetics of Primary Open Angle Glaucoma. *Int J Mol Sci* 16(12), pp. 28886-28911. doi: 10.3390/ijms161226135

Abul-Husn, N. S. et al. 2018. A protein-truncating HSD17B13 variant and protection from chronic liver disease. *New England Journal of Medicine* 378(12), pp. 1096-1106.

Abul-Husn, N. S. et al. 2016. Genetic identification of familial hypercholesterolemia within a single US health care system. *Science* 354(6319), p. aaf7000.

Afek, A., Schipper, J. L., Horton, J., Gordân, R. and Lukatsky, D. B. 2014. Protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci U S A* 111(48), pp. 17140-17145. doi: 10.1073/pnas.1410569111

Aldahmesh, M. A. et al. 2013. Mutations in LRPAP1 are associated with severe myopia in humans. *The American Journal of Human Genetics* 93(2), pp. 313-320.

Alfano, G. et al. 2005. Natural antisense transcripts associated with genes involved in eye development. *Human Molecular Genetics* 14(7), pp. 913-923. doi: 10.1093/hmg/ddi084

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. and Wjst, M. 2001. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 69(5), pp. 936-950. doi: 10.1086/324069

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. and Zondervan, K. T. 2010. Data quality control in genetic case-control association studies. *Nature protocols* 5(9), pp. 1564-1573.

Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* 59(1), pp. 19-35.

Andrew, S. E. et al. 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat Genet* 4(4), pp. 398-403. doi: 10.1038/ng0893-398

Au Eong, K. G., Tay, T. H. and Lim, M. K. 1993. Education and myopia in 110,236 young Singaporean males. *Singapore Med J* 34(6), pp. 489-492.

Baird, P. N. et al. 2020. Myopia. *Nat Rev Dis Primers* 6(1), p. 99. doi: 10.1038/s41572-020-00231-4

Bender, R. 2009. Introduction to the use of regression models in epidemiology. *Cancer Epidemiology*. Springer, pp. 179-195.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1), pp. 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Biscotti, M. A., Olmo, E. and Heslop-Harrison, J. S. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Res* 23(3), pp. 415-420. doi: 10.1007/s10577-015-9499-z

Bolognini, D., Magi, A., Benes, V., Korbel, J. O. and Rausch, T. 2020. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. *Gigascience* 9(10), doi: 10.1093/gigascience/giaa101

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. and Taberlet, P. 2004. How to track and assess genotyping errors in population genetics studies. *Mol Ecol* 13(11), pp. 3261-3273. doi: 10.1111/j.1365-294X.2004.02346.x

Bonvicini, C. et al. 2019. Next generation sequencing analysis in early onset dementia patients. *Journal of Alzheimer's disease* 67(1), pp. 243-256.

Borecki, I. B. and Province, M. A. 2008. Genetic and Genomic Discovery Using Family Studies. *Circulation* 118(10), pp. 1057-1063. doi: doi:10.1161/CIRCULATIONAHA.107.714592

Bremer, D. L., Palmer, E. A., Fellows, R. R., Baker, J. D., Hardy, R. J., Tung, B. and Rogers, G. L. 1998. Strabismus in premature infants in the first year of life. Cryotherapy for Retinopathy of Prematurity Cooperative Group. *Arch Ophthalmol* 116(3), pp. 329-333. doi: 10.1001/archopht.116.3.329

Brown, S. A., Weih, L. M., Fu, C. L., Dimitrov, P., Taylor, H. R. and McCarty, C. A. 2000. Prevalence of amblyopia and associated refractive errors in an adult population in Victoria, Australia. *Ophthalmic Epidemiology* 7(4), pp. 249-258. doi: 10.1076/0928-6586(200012)741-YFT249

Bush, W. S. and Moore, J. H. 2012. Chapter 11: Genome-wide association studies. *PLoS computational biology* 8(12), p. e1002822.

Bycroft, C. et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562(7726), pp. 203-209. doi: 10.1038/s41586-018-0579-z

Cai, X.-B., Shen, S.-R., Chen, D.-F., Zhang, Q. and Jin, Z.-B. 2019. An overview of myopia genetics. *Experimental Eye Research* 188, p. 107778. doi: 10.1016/j.exer.2019.107778

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. and Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4(1), pp. s13742-13015-10047-13748.

Charlesworth, B., Sniegowski, P. and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371(6494), pp. 215-220. doi: 10.1038/371215a0

Chen, D. et al. 2020. Prevalence, incidence and risk factors of strabismus in a Chinese population-based cohort of preschool children: the Nanjing Eye Study. *British Journal of Ophthalmology*, pp. bjophthalmol-2020-316807. doi: 10.1136/bjophthalmol-2020-316807

Chen, J. L., VanEtten, D. M., Fountain, M. A., Yildirim, I. and Disney, M. D. 2017. Structure and Dynamics of RNA Repeat Expansions That Cause Huntington's Disease and Myotonic Dystrophy Type 1. *Biochemistry* 56(27), pp. 3463-3474. doi: 10.1021/acs.biochem.7b00252

Cheng, C. Y. et al. 2013. Nine loci for ocular axial length identified through genome-wide association studies, including shared loci with refractive error. *Am J Hum Genet* 93(2), pp. 264-277. doi: 10.1016/j.ajhg.2013.06.016

Chew, C. K., Foster, P., Hurst, J. A. and Salmon, J. F. 1995. Duane's retraction syndrome associated with chromosome 4q27-31 segment deletion. *American journal of ophthalmology* 119(6), pp. 807-809.

Chew, E., Remaley, N. A., Tamboli, A., Zhao, J., Podgor, M. J. and Klebanoff, M. 1994. Risk factors for esotropia and exotropia. *Arch Ophthalmol* 112(10), pp. 1349-1355. doi: 10.1001/archopht.1994.01090220099030

Chua, S. Y. et al. 2016. Age of onset of myopia predicts risk of high myopia in later childhood in myopic Singapore children. *Ophthalmic Physiol Opt* 36(4), pp. 388-394. doi: 10.1111/opo.12305

Ciner, E., Wojciechowski, R., Ibay, G., Bailey-Wilson, J. E. and Stambolian, D. 2008. Genomewide

scan of ocular refraction in African-American families shows significant linkage to chromosome 7p15. *Genetic epidemiology* 32(5), pp. 454-463. doi: 10.1002/gepi.20318

Clayton, D. G. et al. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics* 37(11), pp. 1243-1246. doi: 10.1038/ng1653

Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H. and Hobbs, H. H. 2006. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine* 354(12), pp. 1264-1272.

Conlon, E. G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N. A. and Manley, J. L. 2016. The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *Elife* 5, doi: 10.7554/eLife.17820

Cotter, S. A. et al. 2011. Risk Factors Associated with Childhood Strabismus: The Multi-Ethnic Pediatric Eye Disease and Baltimore Pediatric Eye Disease Studies. *Ophthalmology* 118(11), pp. 2251-2261. doi: 10.1016/j.ophtha.2011.06.032

Cuellar-Partida, G. et al. 2016. Assessing the Genetic Predisposition of Education on Myopia: A Mendelian Randomization Study. *Genet Epidemiol* 40(1), pp. 66-72. doi: 10.1002/gepi.21936

Cumberland, P. M., Bao, Y., Hysi, P. G., Foster, P. J., Hammond, C. J. and Rahi, J. S. 2015. Frequency and Distribution of Refractive Error in Adult Life: Methodology and Findings of the UK Biobank Study. *PLoS One* 10(10), p. e0139780. doi: 10.1371/journal.pone.0139780

Curtin, B. J. 1985. The myopias. *Basic science and clinical management*, doi: 10.1097/00006982-198600620-00013

Dadd, T., Weale, M. E. and Lewis, C. M. 2009. A critical evaluation of genomic control methods for genetic association studies. *Genetic epidemiology* 33(4), pp. 290-298. doi: https://doi.org/10.1002/gepi.20379

Dashnow, H. et al. 2018. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *bioRxiv*, p. 159228. doi: 10.1101/159228

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. and Pollock, D. D. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics* 7(12), p. e1002384. doi: 10.1371/journal.pgen.1002384

De Roeck, A. et al. 2019. NanoSatellite: accurate characterization of expanded tandem repeat

length and sequence through whole genome long-read sequencing on PromethION. *Genome Biol* 20(1), p. 239. doi: 10.1186/s13059-019-1856-3

De Roeck, A. et al. 2018. An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathol* 135(6), pp. 827-837. doi: 10.1007/s00401-018-1841-z

Devlin, B. and Roeder, K. 1999. Genomic control for association studies. *Biometrics* 55(4), pp. 997-1004. doi: 10.1111/j.0006-341x.1999.00997.x

Dewey, F. E. et al. 2017. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. *New England Journal of Medicine* 377(3), pp. 211-221.

Ding, B. Y., Shih, Y. F., Lin, L. L. K., Hsiao, C. K. and Wang, I. J. 2017. Myopia among schoolchildren in East Asia and Singapore. *Surv Ophthalmol* 62(5), pp. 677-697. doi: 10.1016/j.survophthal.2017.03.006

Dolzhenko, E. et al. 2017. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* 27(11), pp. 1895-1903. doi: 10.1101/gr.225672.117

Doyu, M., Sobue, G., Mukai, E., Kachi, T., Yasuda, T., Mitsuma, T. and Takahashi, A. 1992. Severity of X-linked recessive bulbospinal neuronopathy correlates with size of the tandem CAG repeat in androgen receptor gene. *Ann Neurol* 32(5), pp. 707-710. doi: 10.1002/ana.410320517

Durand, E. Y., Eriksson, N. and McLean, C. Y. 2014. Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. *Molecular Biology and Evolution* 31(8), pp. 2212-2222. doi: 10.1093/molbev/msu151

Engle, E. C. 2006. The genetic basis of complex strabismus. *Pediatr Res* 59(3), pp. 343-348. doi: 10.1203/01.pdr.0000200797.91630.08

Fan, Q. et al. 2012. Genetic Variants on Chromosome 1q41 Influence Ocular Axial Length and High Myopia. *PLOS Genetics* 8(6), p. e1002753. doi: 10.1371/journal.pgen.1002753

Fan, Q. et al. 2014. Education influences the association between genetic variants and refractive error: a meta-analysis of five Singapore studies. *Human Molecular Genetics* 23(2), pp. 546-554.

Fang, L., Liu, Q., Monteys, A. M., Gonzalez-Alegre, P., Davidson, B. L. and Wang, K. 2022. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome biology* 23(1), p. 108. doi: 10.1186/s13059-022-02670-6

Farbrother, J. E., Kirov, G., Owen, M. J. and Guggenheim, J. A. 2004. Family Aggregation of High

Myopia: Estimation of the Sibling Recurrence Risk Ratio. *Investigative Ophthalmology & Visual Science* 45(9), pp. 2873-2878. doi: 10.1167/iovs.03-1155

Fautsch, M. P. et al. 2021. TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. *Progress in retinal and eye research* 81, p. 100883. doi: https://doi.org/10.1016/j.preteyeres.2020.100883

Fearnley, L. G., Bennett, M. F. and Bahlo, M. 2022. Detection of repeat expansions in large next generation DNA and RNA sequencing data without alignment. *Scientific Reports* 12(1), p. 13124. doi: 10.1038/s41598-022-17267-z

Fieß, A., Kölb-Keerl, R., Schuster, A. K., Knuf, M., Kirchhof, B., Muether, P. S. and Bauer, J. 2017. Prevalence and associated factors of strabismus in former preterm and full-term infants between 4 and 10 Years of age. *BMC ophthalmology* 17(1), pp. 1-9.

Figueroa, K. P., Coon, H., Santos, N., Velazquez, L., Mederos, L. A. and Pulst, S. M. 2017. Genetic analysis of age at onset variation in spinocerebellar ataxia type 2. *Neurol Genet* 3(3), p. e155. doi: 10.1212/nxg.0000000000000155

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80(1), pp. 27-38. doi: 10.1093/biomet/80.1.27

Fisher, S. A. et al. 2008. Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nature genetics* 40(6), pp. 710-712. doi: 10.1038/ng.145

Flitcroft, D. I. et al. 2019. IMI - Defining and Classifying Myopia: A Proposed Set of Standards for Clinical and Epidemiologic Studies. *Invest Ophthalmol Vis Sci* 60(3), pp. M20-m30. doi: 10.1167/iovs.18-25957

Forster, P., Hohoff, C., Dunkelmann, B., Schürenkamp, M., Pfeiffer, H., Neuhuber, F. and Brinkmann, B. 2015. Elevated germline mutation rate in teenage fathers. *Proceedings of the Royal Society B: Biological Sciences* 282(1803), p. 20142898.

Fotouhi, A., Etemadi, A., Hashemi, H., Zeraati, H., Bailey-Wilson, J. E. and Mohammad, K. 2007. Familial aggregation of myopia in the Tehran eye study: estimation of the sibling and parent–offspring recurrence risk ratios. *British Journal of Ophthalmology* 91(11), pp. 1440-1444.

Fricke, T. R., Holden, B. A., Wilson, D. A., Schlenther, G., Naidoo, K. S., Resnikoff, S. and Frick, K. D. 2012. Global cost of correcting vision impairment from uncorrected refractive error. *Bull World Health Organ* 90(10), pp. 728-738. doi: 10.2471/blt.12.104034

Fricke, T. R. et al. 2018. Global prevalence of visual impairment associated with myopic macular degeneration and temporal trends from 2000 through 2050: systematic review, meta-analysis and modelling. *British Journal of Ophthalmology* 102(7), pp. 855-862. doi: 10.1136/bjophthalmol-2017-311266

Fujimoto, M., Hangai, M., Suda, K. and Yoshimura, N. 2010. Features associated with foveal retinal detachment in myopic macular retinoschisis. *American journal of ophthalmology* 150(6), pp. 863-870. e861.

Fulton, A. B., Hansen, R. M. and Petersen, R. A. 1982. The Relation of Myopia and Astigmatism in Developing Eyes. *Ophthalmology* 89(4), pp. 298-302. doi: 10.1016/S0161-6420(82)34788-0

Gal, A. et al. 2011. Autosomal-recessive posterior microphthalmos is caused by mutations in PRSS56, a gene encoding a trypsin-like serine protease. *Am J Hum Genet* 88(3), pp. 382-390. doi: 10.1016/j.ajhg.2011.02.006

Galton, F. 1883. *Inquiries into human faculty and its development*. Macmillan.

Gao, X. R., Huang, H. and Kim, H. 2019. Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort. *Human Molecular Genetics* 28(7), pp. 1162-1172.

Gelfand, Y., Hernandez, Y., Loving, J. and Benson, G. 2014. VNTRseek-a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Res* 42(14), pp. 8884-8894. doi: 10.1093/nar/gku642

Gelfand, Y., Rodriguez, A. and Benson, G. 2007. TRDB—the tandem repeats database. *Nucleic acids research* 35(suppl_1), pp. D80-D87.

Gemayel, R., Vinces, M. D., Legendre, M. and Verstrepen, K. J. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* 44, pp. 445-477. doi: 10.1146/annurev-genet-072610-155046

Giesselmann, P. et al. 2019. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat Biotechnol* 37(12), pp. 1478-1481. doi: 10.1038/s41587-019-0293-x

Gordon, R. A. and Donzis, P. B. 1985. Refractive development of the human eye. *Arch Ophthalmol* 103(6), pp. 785-789. doi: 10.1001/archopht.1985.01050060045020

Goss, D., Hampton, M. and Wickham, M. 1988. Selected review on genetic factors in myopia.

Journal of the American Optometric Association 59(11), pp. 875-884.

Graeber, C. P., Hunter, D. G. and Engle, E. C. 2013. The genetic basis of incomitant strabismus: consolidation of the current knowledge of the genetic foundations of disease. *Semin Ophthalmol* 28(5-6), pp. 427-437. doi: 10.3109/08820538.2013.825288

Griffith, J. F., Wilson, R., Cimino, H. C., Patthoff, M., Martin, D. F. and Traboulsi, E. I. 2016. The use of a mobile van for school vision screening: results of 63 841 evaluations. *American journal of ophthalmology* 163, pp. 108-114. e101.

Grosvenor, T., Perrigin, D. M., Perrigin, J. and Maslovitz, B. 1987. Houston Myopia Control Study: a randomized clinical trial. Part II. Final report by the patient care team. *American journal of optometry and physiological optics* 64(7), pp. 482-498.

Guggenheim, J. A. et al. 2022. Whole exome sequence analysis in 51 624 participants identifies novel genes and variants associated with refractive error and myopia. *Human Molecular Genetics* 31(11), pp. 1909-1919. doi: 10.1093/hmg/ddac004

Guggenheim, J. A., Kirov, G. and Hodson, S. A. 2000. The heritability of high myopia: a reanalysis of Goldschmidt's data. *J Med Genet* 37(3), pp. 227-231. doi: 10.1136/jmg.37.3.227

Guo, H. et al. 2014a. SLC39A5 mutations interfering with the BMP/TGF-β pathway in non-syndromic high myopia. *Journal of medical genetics* 51(8), pp. 518-525.

Guo, H. et al. 2014b. SLC39A5 mutations interfering with the BMP/TGF-β pathway in non-syndromic high myopia. *Journal of medical genetics* 51(8), pp. 518-525. doi: 10.1136/jmedgenet-2014-102351

Guo, H. et al. 2015. Mutations of P4HA2 encoding prolyl 4-hydroxylase 2 are associated with nonsyndromic high myopia. *Genetics in Medicine* 17(4), pp. 300-306.

Guo, L. et al. 2016. Prevalence and associated factors of myopia among primary and middle school-aged students: a school-based study in Guangzhou. *Eye (Lond)* 30(6), pp. 796-804. doi: 10.1038/eye.2016.39

Guo, R., Li, Y.-R., He, S., Ou-Yang, L., Sun, Y. and Zhu, Z. 2018. RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics* 34(7), pp. 1099-1107.

Guo, X., Xiao, X., Li, S., Wang, P., Jia, X. and Zhang, Q. 2010. Nonsyndromic High Myopia in a Chinese Family Mapped to MYP1: Linkage Confirmation and Phenotypic Characterization. *Archives of Ophthalmology* 128(11), pp. 1473-1479. doi: 10.1001/archophthalmol.2010.270

Gwiazda, J., Grice, K., Held, R., McLellan, J. and Thorn, F. 2000. Astigmatism and the development of myopia in children. *Vision Research* 40(8), pp. 1019-1026. doi: 10.1016/S0042-6989(99)00237-0

Gwiazda, J. et al. 2007. Factors Associated with High Myopia After 7 Years of Follow-up in the Correction of Myopia Evaluation Trial (COMET) Cohort. *Ophthalmic Epidemiology* 14(4), pp. 230-237. doi: 10.1080/01658100701486459

Gwiazda, J., Thorn, F., Bauer, J. and Held, R. 1993. Emmetropization and the progression of manifest refraction in children followed from infancy to puberty. *Clinical vision sciences* 8(4), pp. 337-344.

Gymrek, M., Golan, D., Rosset, S. and Erlich, Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* 22(6), pp. 1154-1162. doi: 10.1101/gr.135780.111

Gymrek, M. et al. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics* 48(1), pp. 22-29. doi: 10.1038/ng.3461

Hakim, R. B., Stewart, W. F., Canner, J. K. and Tielsch, J. M. 1991. Occupational lead exposure and strabismus in offspring: a case-control study. *Am J Epidemiol* 133(4), pp. 351-356. doi: 10.1093/oxfordjournals.aje.a115888

Hammond, C. J., Andrew, T., Mak, Y. T. and Spector, T. D. 2004. 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* 75(2), pp. 294-304. doi: 10.1086/423148

Hammond, C. J., Snieder, H., Gilbert, C. E. and Spector, T. D. 2001. Genes and environment in refractive error: the twin eye study. *Investigative Ophthalmology & Visual Science* 42(6), pp. 1232-1236.

Hannan, A. J. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics* 19(5), pp. 286-298.

Harman, N. B. 1913. The Education of High Myopes. *Proc R Soc Med* 6(Sect Ophthalmol), pp. 146-163.

Hartwig, A., Gowen, E., Charman, W. N. and Radhakrishnan, H. 2011. Working distance and eye and head movements during near work in myopes and non-myopes. *Clinical and Experimental Optometry* 94(6), pp. 536-544. doi: https://doi.org/10.1111/j.1444-0938.2011.00623.x

Hashemi, H. et al. 2015. The prevalence of strabismus in 7-year-old schoolchildren in Iran. *Strabismus* 23(1), pp. 1-7.

Hatt, S. R., Leske, D. A., Liebermann, L. and Holmes, J. M. 2016. Symptoms in children with intermittent exotropia and their impact on health-related quality of life. *Strabismus* 24(4), pp. 139-145.

He, M. et al. 2015. Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial. *JAMA* 314(11), pp. 1142-1148. doi: 10.1001/jama.2015.10803

He, Z. et al. 2021. Genome-wide analysis of common and rare variants via multiple knockoffs at biobank scale, with an application to Alzheimer disease genetics. *Am J Hum Genet* 108(12), pp. 2336-2353. doi: 10.1016/j.ajhg.2021.10.009

Heidary, G., Ying, G.-S., Maguire, M. G. and Young, T. L. 2005. The association of astigmatism and spherical refractive error in a high myopia cohort. *Optometry and vision science* 82(4), pp. 244-247.

Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A. and Mittelman, D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 41(1), p. e32. doi: 10.1093/nar/gks981

Hirsch, M. J. 1964. Predictability of refraction at age 14 on the basis of testing at age 6--interim report from the OJAI longitudinal study of refraction. *American journal of optometry and archives of American Academy of Optometry* 41, pp. 567-573.

Hitzenberger, C. K. 1991. Optical measurement of the axial eye length by laser Doppler interferometry. *Invest Ophthalmol Vis Sci* 32(3), pp. 616-624.

Holden, B., Sankaridurg, P., Smith, E., Aller, T., Jong, M. and He, M. 2014. Myopia, an underrated global challenge to vision: where the current data takes us on myopia control. *Eye (Lond)* 28(2), pp. 142-146. doi: 10.1038/eye.2013.256

Holden, B. A. et al. 2016. Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050. *Ophthalmology* 123(5), pp. 1036-1042. doi: 10.1016/j.ophtha.2016.01.006

Hosoda, Y. et al. 2018. CCDC102B confers risk of low vision and blindness in high myopia. *Nature Communications* 9(1), p. 1782. doi: 10.1038/s41467-018-03649-3

Hsu, A. R., Mountain, J. L., Wojcicki, A. and Avey, L. 2009. A Pragmatic Consideration of Ethical Issues Relating to Personal Genomics. *The American Journal of Bioethics* 9(6-7), pp. 1-2. doi: 10.1080/15265160902966795

Hu, Y., Ding, X., Guo, X., Chen, Y., Zhang, J. and He, M. 2020. Association of Age at Myopia Onset With Risk of High Myopia in Adulthood in a 12-Year Follow-up of a Chinese Cohort. *JAMA Ophthalmology* 138(11), pp. 1129-1134. doi: 10.1001/jamaophthalmol.2020.3451

Huang, H.-M., Chang, D. S.-T. and Wu, P.-C. 2015. The association between near work activities and myopia in children—a systematic review and meta-analysis. *PLoS One* 10(10), p. e0140419.

Hysi, P. G. et al. 2020. Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia. *Nature genetics* 52(4), pp. 401-407. doi: 10.1038/s41588-020-0599-0

Hysi, P. G. et al. 2010. A genome-wide association study for myopia and refractive error identifies a susceptibility locus at 15q25. *Nature genetics* 42(10), pp. 902-905. doi: 10.1038/ng.664

Igarashi, S. et al. 1992. Strong correlation between the number of CAG repeats in androgen receptor genes and the clinical onset of features of spinal and bulbar muscular atrophy. *Neurology* 42(12), pp. 2300-2302. doi: 10.1212/wnl.42.12.2300

Illarioshkin, S. N. et al. 1994. Trinucleotide repeat length and rate of progression of Huntington's disease. *Ann Neurol* 36(4), pp. 630-635. doi: 10.1002/ana.410360412

Ip, J. M. et al. 2008. Ethnic differences in refraction and ocular biometry in a population-based sample of 11-15-year-old Australian children. *Eye (Lond)* 22(5), pp. 649-656. doi: 10.1038/sj.eye.6702701

Iwase, A., Araie, M., Tomidokoro, A., Yamamoto, T., Shimizu, H. and Kitazawa, Y. 2006. Prevalence and causes of low vision and blindness in a Japanese adult population: the Tajimi Study. *Ophthalmology* 113(8), pp. 1354-1362. doi: 10.1016/j.ophtha.2006.04.022

Jensen, H. 1995. Myopia in teenagers. *Acta Ophthalmologica Scandinavica* 73(5), pp. 389-393. doi: https://doi.org/10.1111/j.1600-0420.1995.tb00294.x

Jiang, D. et al. 2015. Detection of mutations in LRPAP1, CTSH, LEPREL1, ZNF644, SLC39A5, and SCO2 in 298 families with early-onset high myopia by exome sequencing. *Investigative Ophthalmology & Visual Science* 56(1), pp. 339-345.

Jin, Z. B. et al. 2017. Trio-based exome sequencing arrests de novo mutations in early-onset high myopia. *Proc Natl Acad Sci U S A* 114(16), pp. 4219-4224. doi: 10.1073/pnas.1615970114

Johansson, J., Forsgren, L., Sandgren, O., Brice, A., Holmgren, G. and Holmberg, M. 1998. Expanded CAG repeats in Swedish spinocerebellar ataxia type 7 (SCA7) patients: effect of CAG repeat length on the clinical manifestation. *Hum Mol Genet* 7(2), pp. 171-176. doi: 10.1093/hmg/7.2.171

Kawaguchi, Y. et al. 1994. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet* 8(3), pp. 221-228. doi: 10.1038/ng1194-221

Keogh, M. J. and Chinnery, P. F. 2013. Next generation sequencing for neurological diseases: new hope or new hype? *Clin Neurol Neurosurg* 115(7), pp. 948-953. doi: 10.1016/j.clineuro.2012.09.030

Kiefer, A. K., Tung, J. Y., Do, C. B., Hinds, D. A., Mountain, J. L., Francke, U. and Eriksson, N. 2013. Genome-Wide Analysis Points to Roles for Extracellular Matrix Remodeling, the Visual Cycle, and Neuronal Development in Myopia. *PLOS Genetics* 9(2), p. e1003299. doi: 10.1371/journal.pgen.1003299

Klein, A. P., Duggal, P., Lee, K. E., Klein, R., Bailey-Wilson, J. E. and Klein, B. E. 2007. Confirmation of linkage to ocular refraction on chromosome 22q and identification of a novel linkage region on 1q. *Arch Ophthalmol* 125(1), pp. 80-85. doi: 10.1001/archopht.125.1.80

Kleinstein, R. N. et al. 2003. Refractive error and ethnicity in children. *Arch Ophthalmol* 121(8), pp. 1141-1147. doi: 10.1001/archopht.121.8.1141

Klintschar, M., Dauber, E.-M., Ricci, U., Cerri, N., Immel, U.-D., Kleiber, M. and Mayr, W. R. 2004. Haplotype studies support slippage as the mechanism of germline mutations in short tandem repeats. *ELECTROPHORESIS* 25(20), pp. 3344-3348. doi: https://doi.org/10.1002/elps.200406069

Komure, O. et al. 1995. DNA analysis in hereditary dentatorubral-pallidoluysian atrophy: correlation between CAG repeat length and phenotypic variation and the molecular basis of anticipation. *Neurology* 45(1), pp. 143-149. doi: 10.1212/wnl.45.1.143

Kremer, E. J. et al. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* 252(5013), pp. 1711-1714. doi: 10.1126/science.1675488

Kruger, J. M., Mansouri, B. and Cestari, D. M. eds. 2013. *An update on the genetics of comitant strabismus. Seminars in Ophthalmology.* Taylor & Francis.

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. and Fischbeck, K. H. 1991. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352(6330), pp. 77-79. doi: 10.1038/352077a0

Lam, C. Y. et al. 2008a. A genome-wide scan maps a novel high myopia locus to 5p15. *Investigative Ophthalmology & Visual Science* 49(9), pp. 3768-3778.

Lam, D. S., Fan, D. S., Chan, W. M., Tam, B. S., Kwok, A. K., Leung, A. T. and Parsons, H. 2005. Prevalence and characteristics of peripheral retinal degeneration in Chinese adults with high myopia: a cross-sectional prevalence survey. *Optom Vis Sci* 82(4), pp. 235-238. doi: 10.1097/01.opx.0000159359.49457.b4

Lam, D. S. et al. 2008b. The effect of parental history of myopia on children's eye size and growth: results of a longitudinal study. *Investigative Ophthalmology & Visual Science* 49(3), pp. 873-876.

Lam, D. S., Tam, P. O., Fan, D. S., Baum, L., Leung, Y. F. and Pang, C. P. 2003. Familial high myopia linkage to chromosome 18p. *Ophthalmologica* 217(2), pp. 115-118. doi: 10.1159/000068554

Lander, E. S. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822), pp. 860-921. doi: 10.1038/35057062

Lang, J. 1995. Strabismus—Diagnostics, Types of Strabismus, Therapy. *Bern: H Huber*,

Lannes, R., Rizzon, C. and Lerat, E. 2019. Does the Presence of Transposable Elements Impact the Epigenetic Environment of Human Duplicated Genes? *Genes (Basel)* 10(3), doi: 10.3390/genes10030249

Ledford, K. L., Martinez-De Luna, R. I., Theisen, M. A., Rawlins, K. D., Viczian, A. S. and Zuber, M. E. 2017. Distinct cis-acting regions control six6 expression during eye field and optic cup stages of eye formation. *Dev Biol* 426(2), pp. 418-428. doi: 10.1016/j.ydbio.2017.04.003

Lee, K. E., Klein, B. E., Klein, R. and Fine, J. P. 2001. Aggregation of refractive error and 5-year changes in refractive error among families in the Beaver Dam Eye Study. *Archives of Ophthalmology* 119(11), pp. 1679-1685.

Lee, S. and Mackey, D. 2021. 25 Regional Differences in Prevalence of Myopia: Genetic or Environmental Effects? *Advances in Vision Research, Volume III …*,

Li, H. 2014a. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics (Oxford, England)* 30(20), pp. 2843-2851. doi:

10.1093/bioinformatics/btu356

Li, H. 2014b. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 30(20), pp. 2843-2851. doi: 10.1093/bioinformatics/btu356

Li, H. and Homer, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 11(5), pp. 473-483. doi: 10.1093/bib/bbq015

Li, Y., Liu, J. and Qi, P. 2017. The increasing prevalence of myopia in junior high school students in the Haidian District of Beijing, China: a 10-year population-based survey. *BMC Ophthalmol* 17(1), p. 88. doi: 10.1186/s12886-017-0483-6

Li, Y. J. et al. 2011. Genome-wide association studies reveal genetic variants in CTNND2 for high myopia in Singapore Chinese. *Ophthalmology* 118(2), pp. 368-375. doi: 10.1016/j.ophtha.2010.06.016

Liang, C. L. et al. 2004. Impact of family history of high myopia on level and onset of myopia. *Invest Ophthalmol Vis Sci* 45(10), pp. 3446-3452. doi: 10.1167/iovs.03-1058

Lin, Y., Dent, S. Y., Wilson, J. H., Wells, R. D. and Napierala, M. 2010. R loops stimulate genetic instability of CTG.CAG repeats. *Proc Natl Acad Sci U S A* 107(2), pp. 692-697. doi: 10.1073/pnas.0909740107

Liu, Q., Zhang, P., Wang, D., Gu, W. and Wang, K. 2017. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. *Genome medicine* 9, pp. 1-16.

Liu, X. S. et al. 2018. Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. *Cell* 172(5), pp. 979-992.e976. doi: 10.1016/j.cell.2018.01.012

López-Flores, I. and Garrido-Ramos, M. A. 2012. The repetitive DNA content of eukaryotic genomes. *Genome Dyn* 7, pp. 1-28. doi: 10.1159/000337118

Lu, B. et al. 2009. Associations between near work, outdoor activity, and myopia among adolescent students in rural China: the Xichang Pediatric Refractive Error Study report no. 2. *Archives of Ophthalmology* 127(6), pp. 769-775.

Lyhne, N., Sjølie, A. K., Kyvik, K. O. and Green, A. 2001. The importance of genes and environment for ocular refraction and its determiners: a population based study among 20–45 year old twins. *British Journal of Ophthalmology* 85(12), pp. 1470-1476.

Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. and Investigators, G. D. 2013. Recommended joint

and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic epidemiology* 37(6), pp. 539-550.

Ma, J.-H. et al. 2010. Identification of a locus for autosomal dominant high myopia on chromosome 5p13. 3-p15. 1 in a Chinese family. *Molecular vision* 16, p. 2043.

Maconachie, G. D., Gottlob, I. and McLean, R. J. 2013. Risk factors and genetics in common comitant strabismus: a systematic review of the literature. *JAMA Ophthalmol* 131(9), pp. 1179-1186. doi: 10.1001/jamaophthalmol.2013.4001

Mantel, N. 1973. Synthetic retrospective studies and related topics. *Biometrics*, pp. 479-486.

Matamoros, E. et al. 2015. Prevalence of Myopia in France: A Cross-Sectional Analysis. *Medicine (Baltimore)* 94(45), p. e1976. doi: 10.1097/md.0000000000001976

Matsuo, T., Hayashi, M., Fujiwara, H., Yamane, T. and Ohtsuki, H. 2002. Concordance of strabismic phenotypes in monozygotic versus multizygotic twins and other multiple births. *Jpn J Ophthalmol* 46(1), pp. 59-64. doi: 10.1016/s0021-5155(01)00465-8

Matsuo, T., Yamane, T. and Ohtsuki, H. 2001. Heredity versus abnormalities in pregnancy and delivery as risk factors for different types of comitant strabismus. *J Pediatr Ophthalmol Strabismus* 38(2), pp. 78-82.

McKean-Cowdin, R., Cotter, S. A., Tarczy-Hornoch, K., Wen, G., Kim, J., Borchert, M. and Varma, R. 2013. Prevalence of Amblyopia or Strabismus in Asian and Non-Hispanic White Preschool Children: Multi-Ethnic Pediatric Eye Disease Study. *Ophthalmology* 120(10), pp. 2117-2124. doi: 10.1016/j.ophtha.2013.03.001

Meguro, A. et al. 2020. Genome-Wide Association Study in Asians Identifies Novel Loci for High Myopia and Highlights a Nervous System Role in Its Pathogenesis. *Ophthalmology* 127(12), pp. 1612-1624. doi: 10.1016/j.ophtha.2020.05.014

Mehta, C. R., Patel, N. R. and Tsiatis, A. A. 1984. Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics*, pp. 819-825.

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E. and Plohl, M. 2015. Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Research* 23, pp. 583-596.

Mirkin, S. M. 2007. Expandable DNA repeats and human disease. *Nature* 447(7147), pp. 932-940. doi: 10.1038/nature05977

Mitsuhashi, S. et al. 2019. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome biology* 20, pp. 1-17.

Morgan, I. G., Ohno-Matsui, K. and Saw, S.-M. 2012. Myopia. *The Lancet* 379(9827), pp. 1739-1748.

Morgan, I. G. et al. 2021. IMI Risk Factors for Myopia. *Investigative Ophthalmology & Visual Science* 62(5), pp. 3-3. doi: 10.1167/iovs.62.5.3

Mountjoy, E. et al. 2018. Education and myopia: assessing the direction of causality by mendelian randomisation. *BMJ* 361, p. k2022. doi: 10.1136/bmj.k2022

Mukhopadhyay, R. 2009. DNA sequencers: the next generation. *Analytical Chemistry* 81(5), pp. 1736-1740. doi: 10.1021/ac802712u

Mutti, D. O., Mitchell, G. L., Moeschberger, M. L., Jones, L. A. and Zadnik, K. 2002. Parental Myopia, Near Work, School Achievement, and Children's Refractive Error. *Investigative Ophthalmology & Visual Science* 43(12), pp. 3633-3640.

Naidoo, K. S., Raghunandan, A., Mashige, K. P., Govender, P., Holden, B. A., Pokharel, G. P. and Ellwein, L. B. 2003. Refractive error and visual impairment in African children in South Africa. *Invest Ophthalmol Vis Sci* 44(9), pp. 3764-3770. doi: 10.1167/iovs.03-0283

Naiglin, L. et al. 2002. A genome wide scan for familial high myopia suggests a novel locus on chromosome 7q36. *J Med Genet* 39(2), pp. 118-124. doi: 10.1136/jmg.39.2.118

Nair, K. S. et al. 2011. Alteration of the serine protease PRSS56 causes angle-closure glaucoma in mice and posterior microphthalmia in humans and mice. *Nature genetics* 43(6), pp. 579-584. doi: 10.1038/ng.813

Nakanishi, H. et al. 2009. A genome-wide association analysis identified a novel susceptible locus for pathological myopia at 11q24.1. *PLoS Genet* 5(9), p. e1000660. doi: 10.1371/journal.pgen.1000660

Nallasamy, S., Paluru, P. C., Devoto, M., Wasserman, N. F., Zhou, J. and Young, T. L. 2007. Genetic linkage study of high-grade myopia in a Hutterite population from South Dakota. *Molecular vision* 13, p. 229.

Nasrallah, M. P., Cho, G., Simonet, J. C., Putt, M. E., Kitamura, K. and Golden, J. A. 2012. Differential effects of a polyalanine tract expansion in Arx on neural development and gene

expression. *Human molecular genetics* 21(5), pp. 1090-1098. doi: 10.1093/hmg/ddr538

Nelson, C. P. et al. 2017. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature genetics* 49(9), pp. 1385-1391. doi: 10.1038/ng.3913

Nishizaki, R. et al. 2009. New susceptibility locus for high myopia is linked to the uromodulin-like 1 (UMODL1) gene region on chromosome 21q22.3. *Eye* 23(1), pp. 222-229. doi: 10.1038/eye.2008.152

Nurk, S. et al. 2022. The complete sequence of a human genome. *Science* 376(6588), pp. 44-53. doi: 10.1126/science.abj6987

Ohno-Matsui, K. et al. 2015. International photographic classification and grading system for myopic maculopathy. *Am J Ophthalmol* 159(5), pp. 877-883.e877. doi: 10.1016/j.ajo.2015.01.022

Ohno-Matsui, K., Lai, T. Y., Lai, C. C. and Cheung, C. M. 2016. Updates of pathologic myopia. *Prog Retin Eye Res* 52, pp. 156-187. doi: 10.1016/j.preteyeres.2015.12.001

Okui, S., Meguro, A., Takeuchi, M., Yamane, T., Okada, E., Iijima, Y. and Mizuki, N. 2016. Analysis of the association between the LUM rs3759223 variant and high myopia in a Japanese population. *Clinical ophthalmology (Auckland, NZ)* 10, p. 2157.

Orosz, O. et al. 2017. Myopia and Late-Onset Progressive Cone Dystrophy Associate to LVAVA/MVAVA Exon 3 Interchange Haplotypes of Opsin Genes on Chromosome X. *Investigative Ophthalmology & Visual Science* 58(3), pp. 1834-1842. doi: 10.1167/iovs.16-21405

Orozco, G., Barrett, J. C. and Zeggini, E. 2010. Synthetic associations in the context of genome-wide association scan signals. *Hum Mol Genet* 19(R2), pp. R137-144. doi: 10.1093/hmg/ddq368

Orr, A. et al. 2011. Mutations in a novel serine protease PRSS56 in families with nanophthalmos. *Mol Vis* 17, pp. 1850-1861.

Ouyang, J. et al. 2019. CPSF1 mutations are associated with early-onset high myopia and involved in retinal ganglion cell axon projection. *Human Molecular Genetics* 28(12), pp. 1959-1970. doi: 10.1093/hmg/ddz029

Paget, S., Julia, S., Vitezica, Z. G., Soler, V., Malecaze, F. and Calvas, P. 2008. Linkage analysis of high myopia susceptibility locus in 26 families. *Mol Vis* 14, pp. 2566-2574.

Paluru, P. et al. 2003a. New locus for autosomal dominant high myopia maps to the long arm of chromosome 17. *Investigative Ophthalmology & Visual Science* 44(5), pp. 1830-1836.

Paluru, P. et al. 2003b. New locus for autosomal dominant high myopia maps to the long arm of chromosome 17. *Invest Ophthalmol Vis Sci* 44(5), pp. 1830-1836. doi: 10.1167/iovs.02-0697

Paluru, P. C., Nallasamy, S., Devoto, M., Rappaport, E. F. and Young, T. L. 2005. Identification of a novel locus on 2q for autosomal dominant high-grade myopia. *Investigative Ophthalmology & Visual Science* 46(7), pp. 2300-2307.

Pan, C.-W. et al. 2013. Prevalence of Refractive Errors in a Multiethnic Asian Population: The Singapore Epidemiology of Eye Disease Study. *Investigative Ophthalmology & Visual Science* 54(4), pp. 2590-2598. doi: 10.1167/iovs.13-11725

Parikh, V. et al. 2003. A strabismus susceptibility locus on chromosome 7p. *Proceedings of the national academy of sciences* 100(21), pp. 12283-12288.

Park, S. H., Mok, J. and Joo, C.-K. 2013. Absence of an association between lumican promoter variants and high myopia in the Korean population. *Ophthalmic genetics* 34(1-2), pp. 43-47. doi: 10.3109/13816810.2012.736591

Paul, T. O. and Hardage, L. K. 1994. The heritability of strabismus. *Ophthalmic Genet* 15(1), pp. 1-18. doi: 10.3109/13816819409056905

Paulson, H. 2018a. Repeat expansion diseases. *Handbook of clinical neurology* 147, pp. 105-123.

Paulson, H. 2018b. Repeat expansion diseases. *Handbook of clinical neurology* 147, pp. 105-123. doi: 10.1016/b978-0-444-63233-3.00009-9

Penney, J. B., Jr., Vonsattel, J. P., MacDonald, M. E., Gusella, J. F. and Myers, R. H. 1997. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann Neurol* 41(5), pp. 689-692. doi: 10.1002/ana.410410521

Philipp, D. et al. 2022. The relationship between myopia and near work, time outdoors and socioeconomic status in children and adolescents. *BMC Public Health* 22(1), p. 2058. doi: 10.1186/s12889-022-14377-1

Plotnikov, D., Pärssinen, O., Williams, C., Atan, D. and Guggenheim, J. A. 2022. Commonly occurring genetic polymorphisms with a major impact on the risk of nonsyndromic strabismus: replication in a sample from Finland. *Journal of American Association for Pediatric Ophthalmology and Strabismus* 26(1), pp. 12. e11-12. e16.

Plotnikov, D. et al. 2019. A commonly occurring genetic variant within the NPLOC4-TSPAN10-

PDE6G gene cluster is associated with the risk of strabismus. *Hum Genet* 138(7), pp. 723-737. doi: 10.1007/s00439-019-02022-8

Podgor, M. J., Remaley, N. A. and Chew, E. 1996. Associations between siblings for esotropia and exotropia. *Archives of Ophthalmology* 114(6), pp. 739-744.

Ponsonby, A. L., Brown, S. A., Kearns, L. S., MacKinnon, J. R., Scotter, L. W., Cochrane, J. A. and Mackey, D. A. 2007. The association between maternal smoking in pregnancy, other early life characteristics and childhood vision: the Twins Eye Study in Tasmania. *Ophthalmic Epidemiol* 14(6), pp. 351-359. doi: 10.1080/01658100701486467

Pottier, C. et al. 2015. Whole-genome sequencing reveals important role for TBK1 and OPTN mutations in frontotemporal lobar degeneration without motor neuron disease. *Acta neuropathologica* 130, pp. 77-92.

Pozarickij, A., Williams, C., Hysi, P. G., Guggenheim, J. A., Eye, U. K. B. and Vision, C. 2019. Quantile regression analysis reveals widespread evidence for gene-environment or gene-gene interactions in myopia development. *Communications Biology* 2, pp. 167-167. doi: 10.1038/s42003-019-0387-5

Pritchard, J. K. and Rosenberg, N. A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65(1), pp. 220-228. doi: 10.1086/302449

Pumpernik, D., Oblak, B. and Borštnik, B. 2008. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Molecular Genetics and Genomics* 279(1), pp. 53-61.

Qin, X.-J., Margrain, T. H., To, C. H., Bromham, N. and Guggenheim, J. A. 2005. Anisometropia is independently associated with both spherical and cylindrical ametropia. *Investigative Ophthalmology & Visual Science* 46(11), pp. 4024-4031.

Quilez, J. et al. 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic acids research* 44(8), pp. 3750-3762. doi: 10.1093/nar/gkw219

R Core Team (2020). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ratnamala, U. et al. 2011. Refinement of the X-linked Nonsyndromic High-Grade Myopia Locus MYP1 on Xq28 and Exclusion of 13 Known Positional Candidate Genes by Direct Sequencing. *Investigative Ophthalmology & Visual Science* 52(9), pp. 6814-6819. doi: 10.1167/iovs.10-6815

Reich, D. E. and Goldstein, D. B. 2001. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 20(1), pp. 4-16.

Rim, T. H., Kim, S. H., Lim, K. H., Choi, M., Kim, H. Y. and Baek, S. H. 2016. Refractive Errors in Koreans: The Korea National Health and Nutrition Examination Survey 2008-2012. *Korean J Ophthalmol* 30(3), pp. 214-224. doi: 10.3341/kjo.2016.30.3.214

Risch, N. and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* 273(5281), pp. 1516-1517. doi: 10.1126/science.273.5281.1516

Robaei, D., Kifley, A., Gole, G. A. and Mitchell, P. 2006a. The impact of modest prematurity on visual function at age 6 years: findings from a population-based study. *Arch Ophthalmol* 124(6), pp. 871-877. doi: 10.1001/archopht.124.6.871

Robaei, D., Rose, K. A., Kifley, A., Cosstick, M., Ip, J. M. and Mitchell, P. 2006b. Factors Associated with Childhood Strabismus: Findings from a Population-Based Study. *Ophthalmology* 113(7), pp. 1146-1153. doi: 10.1016/j.ophtha.2006.02.019

Roeder, K., Bacanu, S. A., Wasserman, L. and Devlin, B. 2006. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 78(2), pp. 243-252. doi: 10.1086/500026

Roeder, K. and Wasserman, L. 2009. Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics* 24(4), p. 398.

Rose, K. A., Morgan, I. G., Ip, J., Kifley, A., Huynh, S., Smith, W. and Mitchell, P. 2008. Outdoor activity reduces the prevalence of myopia in children. *Ophthalmology* 115(8), pp. 1279-1285. doi: 10.1016/j.ophtha.2007.12.019

Rosenblatt, A., Kumar, B. V., Mo, A., Welsh, C. S., Margolis, R. L. and Ross, C. A. 2012. Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov Disord* 27(2), pp. 272-276. doi: 10.1002/mds.24024

Rothenburg, S., Koch-Nolte, F., Rich, A. and Haag, F. 2001. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc Natl Acad Sci U S A* 98(16), pp. 8985-8990. doi: 10.1073/pnas.121176998

Rubin, D., Dudoit, S. and Van der Laan, M. 2006. A method to increase the power of multiple

testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology* 5(1),

Salicioni, A. M., Xi, M., Vanderveer, L. A., Balsara, B., Testa, J. R., Dunbrack, R. L., Jr. and Godwin, A. K. 2000. Identification and structural analysis of human RBM8A and RBM8B: two highly conserved RNA-binding motif proteins that interact with OVCA1, a candidate tumor suppressor. *Genomics* 69(1), pp. 54-62. doi: 10.1006/geno.2000.6315

Sanfilippo, P. G. et al. 2012. Heritability of strabismus: genetic influence is specific to eso-deviation and independent of refractive error. *Twin Research and Human Genetics* 15(5), pp. 624-630.

Satterfield, D., Keltner, J. L. and Morrison, T. L. 1993. Psychosocial Aspects of Strabismus Study. *Archives of Ophthalmology* 111(8), pp. 1100-1105. doi: 10.1001/archopht.1993.01090080096024

Saw, S.-M. et al. 2002. Nearwork in Early-Onset Myopia. *Investigative Ophthalmology & Visual Science* 43(2), pp. 332-339.

Saw, S. M., Gazzard, G., Shih-Yen, E. C. and Chua, W. H. 2005. Myopia and associated pathological complications. *Ophthalmic and Physiological Optics* 25(5), pp. 381-391.

Schmid, G. F., Papastergiou, G. I., Nickla, D. L., Riva, C. E., Lin, T., Stone, R. A. and Laties, A. M. 1996. Validation of laser Doppler interferometric measurements in vivo of axial eye length and thickness of fundus layers in chicks. *Curr Eye Res* 15(6), pp. 691-696. doi: 10.3109/02713689609008911

Schöls, L., Vieira-Saecker, A. M., Schöls, S., Przuntek, H., Epplen, J. T. and Riess, O. 1995. Trinucleotide expansion within the MJD1 gene presents clinically as spinocerebellar ataxia and occurs most frequently in German SCA patients. *Hum Mol Genet* 4(6), pp. 1001-1005. doi: 10.1093/hmg/4.6.1001

Schuster, A. K., Elflein, H. M., Pokora, R. and Urschitz, M. S. 2017. Kindlicher Strabismus in Deutschland: Prävalenz und Risikogruppen. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* 60(8), pp. 849-855. doi: 10.1007/s00103-017-2578-x

Schwartz, M., Haim, M. and Skarsholm, D. 1990. X-linked myopia: Bornholm eye disease. Linkage to DNA markers on the distal part of Xq. *Clin Genet* 38(4), pp. 281-286.

Scott, R. A. et al. 2016. A genomic approach to therapeutic target validation identifies a glucose-lowering GLP1R variant protective for coronary heart disease. *Science translational medicine* 8(341), pp. 341ra376-341ra376.

Shaaban, S. et al. 2018. Genome-Wide Association Study Identifies a Susceptibility Locus for Comitant Esotropia and Suggests a Parent-of-Origin Effect. *Invest Ophthalmol Vis Sci* 59(10), pp. 4054-4064. doi: 10.1167/iovs.18-24082

Shaaban, S. et al. 2009. Chromosomes 4q28.3 and 7q31.2 as new susceptibility loci for comitant strabismus. *Invest Ophthalmol Vis Sci* 50(2), pp. 654-661. doi: 10.1167/iovs.08-2437

Shah, R. L., Guggenheim, J. A. and Consortium, U. K. B. E. V. 2018. Genome-wide association studies for corneal and refractive astigmatism in UK Biobank demonstrate a shared role for myopia susceptibility loci. *Human genetics* 137(11-12), pp. 881-896. doi: 10.1007/s00439-018-1942-8

Shi, Y. et al. 2011a. Exome sequencing identifies ZNF644 mutations in high myopia. *PLOS Genetics* 7(6), p. e1002084.

Shi, Y. et al. 2011b. Genetic variants at 13q12. 12 are associated with high myopia in the Han Chinese population. *The American Journal of Human Genetics* 88(6), pp. 805-813.

Silverberg, M. S. et al. 2009. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nature genetics* 41(2), pp. 216-220. doi: 10.1038/ng.275

Sim, B., Yap, G.-H. and Chia, A. 2014. Functional and psychosocial impact of strabismus on Singaporean children. *Journal of American Association for Pediatric Ophthalmology and Strabismus* 18(2), pp. 178-182. doi: 10.1016/j.jaapos.2013.11.013

Simpson, C. L. et al. 2014. Genome-wide meta-analysis of myopia and hyperopia provides evidence for replication of 11 loci. *PLoS One* 9(9), pp. e107110-e107110. doi: 10.1371/journal.pone.0107110

Snell, R. G. et al. 1993. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat Genet* 4(4), pp. 393-397. doi: 10.1038/ng0893-393

Solouki, A. M. et al. 2010. A genome-wide association study identifies a susceptibility locus for refractive errors and myopia at 15q14. *Nature genetics* 42(10), pp. 897-901. doi: 10.1038/ng.663

Stambolian, D. 2013. Genetic susceptibility and mechanisms for refractive error. *Clin Genet* 84(2), pp. 102-108. doi: 10.1111/cge.12180

Stambolian, D. et al. 2004. Genomewide linkage scan for myopia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 22q12. *The American*

Journal of Human Genetics 75(3), pp. 448-459.

Stine, O. C., Pleasant, N., Franz, M. L., Abbott, M. H., Folstein, S. E. and Ross, C. A. 1993. Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. *Hum Mol Genet* 2(10), pp. 1547-1549. doi: 10.1093/hmg/2.10.1547

Sulovari, A. et al. 2019. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proceedings of the national academy of sciences* 116(46), pp. 23243-23253.

Sun, W. et al. 2015. Exome Sequencing on 298 Probands With Early-Onset High Myopia: Approximately One-Fourth Show Potential Pathogenic Mutations in RetNet Genes. *Invest Ophthalmol Vis Sci* 56(13), pp. 8365-8372. doi: 10.1167/iovs.15-17555

Sundin, O. H. et al. 2005. Extreme hyperopia is the result of null mutations in MFRP, which encodes a Frizzled-related protein. *Proc Natl Acad Sci U S A* 102(27), pp. 9553-9558. doi: 10.1073/pnas.0501451102

Tang, H. et al. 2017. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am J Hum Genet* 101(5), pp. 700-715. doi: 10.1016/j.ajhg.2017.09.013

Tang, S. M. et al. 2016. Refractive errors and concomitant strabismus: a systematic review and meta-analysis. *Scientific Reports* 6(1), pp. 1-9.

Tang, S. M., Rong, S. S., Young, A. L., Tam, P. O. S., Pang, C. P. and Chen, L. J. 2014. PAX6 Gene Associated with High Myopia: A Meta-analysis. *Optometry and Vision Science* 91(4),

Tang, W. C., Yap, M. K. and Yip, S. P. 2008. A review of current approaches to identifying human genes involved in myopia. *Clinical and Experimental Optometry* 91(1), pp. 4-22. doi: https://doi.org/10.1111/j.1444-0938.2007.00181.x

Tankard, R. M., Bennett, M. F., Degorski, P., Delatycki, M. B., Lockhart, P. J. and Bahlo, M. 2018. Detecting tandem repeat expansions in cohorts sequenced with short-read sequencing data. *bioRxiv*, p. 157792. doi: 10.1101/157792

Tautz, D. and Schlötterer, C. 1994. Simple sequences. *Current Opinion in Genetics & Development* 4(6), pp. 832-837.

Tedja, M. S. et al. 2019. IMI – Myopia Genetics Report. *Investigative Ophthalmology & Visual Science* 60(3), pp. M89-M105. doi: 10.1167/iovs.18-25965

Tedja, M. S. et al. 2018a. Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nature genetics* 50(6), pp. 834-848. doi: 10.1038/s41588-018-0127-7

Tedja, M. S. et al. 2018b. Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nature genetics* 50(6), pp. 834-848. doi: 10.1038/s41588-018-0127-7

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

Tideman, J. W. L. et al. 2021. Evaluation of Shared Genetic Susceptibility to High and Low Myopia and Hyperopia. *JAMA Ophthalmol* 139(6), pp. 601-609. doi: 10.1001/jamaophthalmol.2021.0497

Tkatchenko, A. V. et al. 2015. APLP2 regulates refractive error and myopia development in mice and humans. *PLOS Genetics* 11(8), p. e1005432.

Tran-Viet, K.-N. et al. 2012. Study of a US cohort supports the role of ZNF644 and high-grade myopia susceptibility. *Molecular vision* 18, p. 937.

Tran-Viet, K.-N. et al. 2013. Mutations in SCO2 are associated with autosomal-dominant high-grade myopia. *The American Journal of Human Genetics* 92(5), pp. 820-826.

Treangen, T. J. and Salzberg, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 13(1), pp. 36-46. doi: 10.1038/nrg3117

UK Biobank 2022a. *Category 170*. Available at: https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=170#:~:text=The%20first%20tranche%20of%20UK, made%20available%20in%20October%202020. [Accessed: Sep 30].

UK Biobank 2022b. *Resource 3803*. Available at: https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=3803 [Accessed: Sep 30].

Usdin, K., House, N. C. and Freudenreich, C. H. 2015. Repeat instability during DNA repair: Insights from model systems. *Critical reviews in biochemistry and molecular biology* 50(2), pp. 142-167.

van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G. and Boomsma, D. I. 2012. The continuing value of twin studies in the omics era. *Nature Reviews Genetics* 13(9), pp. 640-653. doi: 10.1038/nrg3243

Verhoeven, V. J., Buitendijk, G. H., Rivadeneira, F., Uitterlinden, A. G., Vingerling, J. R., Hofman, A. and Klaver, C. C. 2013a. Education influences the role of genetics in myopia. *European journal of epidemiology* 28(12), pp. 973-980.

Verhoeven, V. J. et al. 2013b. Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nature genetics* 45(3), pp. 314-318. doi: 10.1038/ng.2554

Verhoeven, V. J., Wong, K. T., Buitendijk, G. H., Hofman, A., Vingerling, J. R. and Klaver, C. C. 2015. Visual consequences of refractive errors in the general population. *Ophthalmology* 122(1), pp. 101-109. doi: 10.1016/j.ophtha.2014.07.030

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101(1), pp. 5-22. doi: 10.1016/j.ajhg.2017.06.005

Vitale, S., Sperduto, R. D. and Ferris, F. L. 2009. Increased prevalence of myopia in the United States between 1971-1972 and 1999-2004. *Archives of Ophthalmology* 127(12), pp. 1632-1639.

von Bubnoff, A. 2008. Next-generation sequencing: the race is on. *Cell* 132(5), pp. 721-723. doi: 10.1016/j.cell.2008.02.028

Wang, B. et al. 2017. A novel potentially causative variant of NDUFAF7 revealed by mutation screening in a Chinese family with pathologic myopia. *Investigative Ophthalmology & Visual Science* 58(10), pp. 4182-4192.

Wang, X. 2014. Firth logistic regression for rare variant association tests. Frontiers Media SA.

Wasserman, L. A. and Roeder, K. 2006. Weighted Hypothesis Testing. *arXiv: Statistics Theory*,

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M. and Erlich, Y. 2017. Genome-wide profiling of heritable and de novo STR variations. *Nature methods* 14(6), pp. 590-592. doi: 10.1038/nmeth.4267

Williams, K. L. et al. 2016. CCNF mutations in amyotrophic lateral sclerosis and frontotemporal dementia. *Nature Communications* 7(1), p. 11253.

Williams, K. M. et al. 2015. Prevalence of refractive error in Europe: the European eye epidemiology (E3) Consortium. *European journal of epidemiology* 30(4), pp. 305-315.

Wilmer, J. B. and Backus, B. T. 2009. Genetic and environmental contributions to strabismus and

phoria: evidence from twins. *Vision Res* 49(20), pp. 2485-2493. doi: 10.1016/j.visres.2009.08.006

Wojciechowski, R., Congdon, N., Bowie, H., Munoz, B., Gilbert, D. and West, S. K. 2005. Heritability of refractive error and familial aggregation of myopia in an elderly American population. *Investigative Ophthalmology & Visual Science* 46(5), pp. 1588-1592.

Wojciechowski, R., Moy, C., Ciner, E., Ibay, G., Reider, L., Bailey-Wilson, J. E. and Stambolian, D. 2006. Genomewide scan in Ashkenazi Jewish families demonstrates evidence of linkage of ocular refraction to a QTL on chromosome 1p36. *Human genetics* 119(4), pp. 389-399.

Wollstein, A. et al. 2017. Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour. *Scientific Reports* 7(1), pp. 1-11.

Wong, T., Foster, P., Johnson, G. and Seah, S. 2002. Education, socioeconomic status, and ocular dimensions in Chinese adults: the Tanjong Pagar Survey. *British Journal of Ophthalmology* 86(9), pp. 963-968.

Woodman, E. C., Read, S. A., Collins, M. J., Hegarty, K. J., Priddle, S. B., Smith, J. M. and Perro, J. V. 2011. Axial elongation following prolonged near work in myopes and emmetropes. *Br J Ophthalmol* 95(5), pp. 652-656. doi: 10.1136/bjo.2010.180323

Wren, J. D. et al. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *The American Journal of Human Genetics* 67(2), pp. 345-356.

Wu, L., Sun, X., Zhou, X. and Weng, C. 2011. Causes and 3-year-incidence of blindness in Jing-An District, Shanghai, China 2001-2009. *BMC Ophthalmol* 11, p. 10. doi: 10.1186/1471-2415-11-10

Wu, L. J. et al. 2015. Prevalence and associated factors of myopia in high-school students in Beijing. *PLoS One* 10(3), p. e0120764.

Xiang, F., He, M., Zeng, Y., Mai, J., Rose, K. A. and Morgan, I. G. 2013. Increases in the prevalence of reduced visual acuity and myopia in Chinese children in Guangzhou over the past 20 years. *Eye (Lond)* 27(12), pp. 1353-1358. doi: 10.1038/eye.2013.194

Xiao, X., Li, S., Jia, X., Guo, X. and Zhang, Q. 2016. X-linked heterozygous mutations in ARR3 cause female-limited early onset high myopia. *Mol Vis* 22, pp. 1257-1266.

Yang, Z., Xiao, X., Li, S. and Zhang, Q. 2009. Clinical and linkage study on a consanguineous Chinese family with autosomal recessive high myopia. *Molecular vision* 15, p. 312.

Ye, X. C., Pegado, V., Patel, M. S. and Wasserman, W. W. 2014. Strabismus genetics across a spectrum of eye misalignment disorders. *Clin Genet* 86(2), pp. 103-111. doi: 10.1111/cge.12367

Young, T. L. et al. 1998a. A second locus for familial high myopia maps to chromosome 12q. *Am J Hum Genet* 63(5), pp. 1419-1424. doi: 10.1086/302111

Young, T. L. et al. 1998b. Evidence that a locus for familial high myopia maps to chromosome 18p. *Am J Hum Genet* 63(1), pp. 109-119. doi: 10.1086/301907

Zhang, Q., Guo, X., Xiao, X., Jia, X., Li, S. and Hejtmancik, J. 2006. Novel locus for X linked recessive high myopia maps to Xq23–q25 but outside MYP1. *Journal of medical genetics* 43(5), pp. e20-e20.

Zhang, Q., Guo, X., Xiao, X., Jia, X., Li, S. and Hejtmancik, J. F. 2005. A new locus for autosomal dominant high myopia maps to 4q22-q27 between D4S1578 and D4S1612. *Mol Vis* 11(64-65), pp. 554-560.

Zhang, Q., Li, S., Xiao, X., Jia, X. and Guo, X. 2007. Confirmation of a genetic locus for X-linked recessive high myopia outside MYP1. *Journal of human genetics* 52(5), pp. 469-472.

Zhang, X. J. et al. 2021. Prevalence of strabismus and its risk factors among school aged children: The Hong Kong Children Eye Study. *Scientific Reports* 11(1), pp. 1-7.

Zhao, F. et al. 2013. Exome sequencing reveals CCDC111 mutation associated with high myopia. *Human genetics* 132(8), pp. 913-921.

Zhou, W. et al. 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics* 50(9), pp. 1335-1341. doi: 10.1038/s41588-018-0184-y