

東京大学大学院新領域創成科学研究科

先端生命科学専攻

2021 年度

修士論文

Polygenic risk score revealing heterogeneity of myopia associated variants in European

2021 年 2 月 12 日提出

指導教員 中山 一大 准教授

夏 添

Xia Tian

Contents

1	Introduction	2
1.1	Myopia	2
1.2	The rising prevalence of myopia and its consequences	3
1.3	Genetic and environmental impact of myopia epidemic.....	3
1.4	Light environment and myopia risk	5
1.5	Polygenic disease and polygenic risk score	6
1.6	Evolutionary mismatch and ancestry environments.....	7
1.7	The purpose of this study	8
2	Subjects and Methods.....	9
2.1	Subjects of this study	9
2.2	Quality control (QC)	10
2.3	Myopia risk prediction	11
2.4	Ancestry proportion of myopia associated variants	12
3	Results	13
3.1	Polygenic risk scores.....	13
3.2	Best-fit P value threshold estimation	13
3.3	Ancestry proportion of myopia associated variants	14
3.4	Comparison of scored variants.....	14
3.5	Identifying the unique variants.....	16
4	Discussion	17
4.1	The heterogeneity of myopia associated-variants	17
4.2	Selective pressure and sunshine duration hypothesis.....	18
4.3	Indication from the results	19
4.4	Challenge of this study	20
5	Acknowledgements	23
6	References	24
7	Tables	33
8	Figures	38

1 Introduction

1.1 Myopia

Refractive error occurs when light could not focus correctly on the retina; it's often caused by the distortion of the eyeball or cornea. Myopia, also known as near-sightedness or short-sightedness, is the primary form of refractive error (Bourne RR et al. 2013). In the myopic eye, light focus in front of the retina results in blurry vision for distant objects and normal vision for near objects.

There are various forms of myopia by clinical appearance, and some temporally occur due to environmental factors or other health problems. Such as pseudo myopia, which is caused by spasm of the ciliary muscle preventing the eye from focusing in the distance, the clear distance vision is often restored after the eyes getting rest (Chan R et al. 2002). Degenerative myopia, also known as pathological or progressive myopia, is characterized by marked fundus changes, such as posterior staphyloma, and is associated with a high refractive error and subnormal visual acuity after correction (Cline D, 1997). Simple myopia is determined when an eye is too long for its optical power or, less commonly, too optically powerful for its axial length. Simple myopia is generally more than -4 to -6 diopters and much more common than the other types of myopia (Goss DA et al. 1997).

Myopia is diagnosed by the refractive status of each eye. The degree of myopia is described in terms of the power of the ideal correction, which is measured in diopters (Grosvenor T. 1987). The definitions of myopia and high myopia vary across previous studies. A commonly used standard is a spherical equivalent of -0.50 diopters or less for myopia and a spherical equivalent of -5.00 diopters or less for high myopia (Holden BA et al. 2016).

Syndromic and pathological high myopia has a fundamental genetic difference from simple myopia (Morgan IG et al. 2018), and all forms of high myopia accounted for a relatively small proportion in the myopia prevalence. In 2010, global myopia and high myopia prevalence were 22.9% and 2.7% (Holden BA et al. 2016), simple myopia was the major form. This study focuses on simple myopia for its universal susceptibility.

1.2 The rising prevalence of myopia and its consequences

Epidemiological studies observed a global trend of rising prevalence of myopia during the last century, regardless of age, gender, and ethnicity (Morgan I et al. 2005; Williams KM et al. 2015; Holden BA et al. 2016). Early records trace back to European cohorts born in 1910s, the prevalence rose from less than 15% to 37% in 2020, and East Asian records of Hongkong, Taiwan, Singapore and South Korea started from 10–30% in 1940s to over 80% in 2010s (Dolgin E. 2015). Although there are region or ethnicity gaps of myopia prevalence, the historical changes in African Americans were similar to European Americans, suggesting that the region gaps or ethnicity gaps will eventually close (Vitale S et al. 2009). In a systematic review and meta-analysis, the global myopia prevalence estimated to be 22.9% in 2000s and 49.8% in 2050 (Holden BA et al. 2016).

Myopia is associated with an increasing risk of several sight-threatening diseases, including glaucoma (open-angle), cataract, retinal detachment and myopic maculopathy or myopia macular degeneration. Even under appropriate refractive correction, the odd ratio of these conditions drastically increased when the myopia degree gets higher (Saw SM et al. 2005). In 2010, worldwide leading causes for 32.4 million blind people were cataract (33%), uncorrected refractive error (21%), and macular degeneration (7%); and for 191 million people with moderate and severe vision impairment were uncorrected refractive error (53%), cataract (18%), and macular degeneration (3%, Bourne RR et al. 2013). The economic burden of uncorrected distance refractive error, largely caused by myopia, was estimated to be US\$202 billion per annum (Smith TST et al. 2009).

1.3 Genetic and environmental impact of the myopia epidemic

Knowledge of myopia has been accumulated over a century through epidemiological studies, and previous studies managed to reveal some direct risk factors, including extensive near work, greater time spent indoors, urbanization, the higher level of modernized education, and a family history of myopia (Foster PJ et al.

2014).

On the other hand, the development of genome-wide association studies (GWAS) promoted the rise of myopia genetic studies since 2009. GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes, Leslie R et al. 2014) included ten myopia GWAS results published during 2009 and 2013, while six of them were focused on pathological and high myopia. Besides, some best results only explained 3.6% of the variance of the refractive error by the identified genetic variants (Verhoeven et al. 2013), presuming a high missing heritability in myopia (Tedja MS et al. 2018). Heritability quantifies the proportion of phenotypic variation in a trait that is due to genetic factors, and the genetic risk prediction accuracy positively associated with both heritability and discovery sample size (Wray NR, et al. 2013). Thus, sample size limited myopia genetic studies. With more and more efforts taken into myopia genetic studies worldwide, from late 2013 to 2020, the discovery sample size has been risen from about 45 thousand to 276 thousand (Verhoeven et al. 2013; Hysi PG et al. 2020), and eventually associated loci explained 18.4% of myopia variance. Furthermore, the full refractive error heritability estimated to be account for 31.1% of the myopia variance and associated with 13,808 polymorphic variants (Hysi PG et al. 2020).

Although it has been widely accepted that myopia is caused by a combination of genetic and environmental factors, the evidence is strong that environmental factors have played a significant role in the current epidemic of myopia. Previous studies debated about to what extent genetic factors determine the variability of refractive error (Hysi PG et al. 2014), a series of data of Chinese, Indian, and Malay myopia prevalence in Singapore used as a controlled model in this argument (Morgan IG et al. 2018). Later studies proposing that the prevalence gap among the 3 cohorts cannot be explained by their genetic background, the data presenting Chinese are more myopic than Indians and Malays. Still, all groups are more myopic than in other parts of the world, and this suggests that it is the environmental exposure in Singapore rather than genetic background dedicated to the prevalence of myopia (Morgan IG et al. 2012). Following

data supported the idea that when the environmental variation is highly reduced, the genetic difference could account for more in appearance (Morgan IG et al. 2018). This also in accordance with the barely 30% thorough phenotypic variance explained proportion.

1.4 Light environment and myopia risk

The crucial question is to what extend the gene and environment interaction impacts the myopia prevalence. In clinical practice, myopia management and preventing precautions includes atropine 0.01% eye drops, wearing orthokeratology lenses and 2 hours per day sunlight. These clinical data also provided evidence that myopia risk correlated to daylight exposure, a person with little exposure to daylight has a fivefold risk of developing myopia, which can rise as high as a 16-fold risk if that person also performs close-up work (Lagrèze WA et al. 2017). In a meta-analysis and systematic review (Xiong et al. 2017), 13 cross-sectional studies were pooled after their conversion into a standardized effect estimate, yielding a final odd ratio (OR) of 0.964 for myopia per additional hour of time spent outdoors per week.

Time spent outdoors positively associated with the prevention of myopia development is not the whole picture. Moreover, illuminance intensity is essential. In a chicken deprivation myopia model, high illuminance levels (15,000 lux) reduce the rate of compensation for negative lenses and enhance the rate for positive lenses, in other words, strong lightening showed a protective effect in developing myopia (Ashby RS et al. 2010). A previous study (Landis EG et al. 2018) also observed association between light intensity and myopia in children, in which myopic children received significantly less scotopic light (<1 lux, $P=0.024$) and less outdoor photopic light (>1000 lux) than nonmyopic children ($P<0.001$). In myopic children, more myopic refractive errors were correlated with increased time in mesopic light (1–30 lux, $R=-0.46$, $P=0.002$).

The outdoor activity is essentially implying direct sunshine, even the brightest indoor spaces are dim compared to the outdoors in daylight. The intensity levels of typical indoor lighting are 100–500 lux, while outdoor lighting ranges from 1000 to

over 100,000 lux depending on atmospheric conditions (National Optical Astronomy Observatory). In South Korean children (Choo HG et al. 2019), a negative but not statistically significant association between sunshine duration and myopia prevalence ($P=0.064$) was observed. Besides, solar radiation and sunshine duration were significantly associated with mean spherical equivalent ($P=0.001$ and 0.014 respectively).

1.5 Polygenic disease and polygenic risk score

From the genetic aspects, myopia is categorized as polygenic disease, or multifactorial disorder, the polygenic disease is associated with the effects of multiple genes in combination with lifestyles and environmental factors. On the contrary, monogenic disease or Mendelian disease is controlled by a single locus in an inheritance pattern, in this case, a mutation in the very single gene is responsible for disease (Chial H. 2008). In general, polygenic diseases are more common than monogenic diseases, but genetic analysis of polygenic traits are way more challenging because of gene-gene and gene-environment interactions, genetic heterogeneity, low penetrance, and limited statistical power. From 1980s to 2000s, over 1,700 Mendelian traits genes had been identified while less than 10 polygenic traits genes been identified (Glazier AM et al. 2002).

Polygenic risk score (PRS) is an estimate of an individual's genetic susceptibility to a trait or disease, calculated according to their genotype profile and relevant genome-wide association study (GWAS) data (Choi et al. 2020). Its first successful appliance in humans was on schizophrenia in 2009 (Purcell SM et al. 2009), this study was also the first to use the term *polygenic score* for a prediction drawn from a linear combination of single-nucleotide polymorphism (SNP) genotypes. It came out the idea that even no single locus is statistically significant, but the aggregation of these variants is still predictive to the risk of having the disease. Though hard to understand each locus, it is still a promising tool (Martin AR et al. 2019).

Despite that PRS typically explains only a small fraction of trait variance, its

correlation with genetic susceptibility and phenotypic variation makes it versatile in biomedical research. As GWAS sample sizes increase and PRS becomes more powerful, PRS is set to play a key role in research and stratified medicine (Choi SW et al. 2020). Essentially, PRS is still incomplete for clinical appliance due to low power for limited discovery data sample size and ethnic disparities. 23andMe, Inc., a private personal genomics and biotechnology company, has been proposing incorporate genetic risk score into clinical practice (Knowles JW et al. 2018). At the 87th National Advisory Council for Human Genome Research in Sep. 2019, the National Institutes of Health (NIH) recommended and called for further investment in expanding the population diversity in the present database to enhance the prediction accuracy for the potential utilities of PRS in clinical practice.

1.6 Evolutionary mismatch and ancestry environments

Since the human gene pools cannot change rapidly within a century, the drastic rise of prevalence of myopia was driven by a mismatch to the modern environments (Morgan IG et al. 2018). In evolutionary medicine, myopia is also defined as an evolutionary mismatch (Long E. 2018). Humans are more adapted to the original hunter-gatherer life circumstances, and the human genome is shaped by adaption to ancestral environments. Though not been emphasized with the title evolutionary medicine, the present refractive error management and prediction method of spending more time outdoors is essentially an evolutionary treatment. Further trails had been conducted in Guangzhou (Zhou Z et al. 2017), another myogenic site similar to Singapore (Morgan IG et al. 2018), by designing a Bright Classroom prototype to simulate outdoor light conditions while limiting the upper light intensity for children's comfort reading.

The ancestral environment variations also forged genome and phenotype variations such as skin color (Deng L et al. 2017) and lactose tolerance (Bayoumi et al. 2016). It is not hard to associate whether light intensity impacted myopia-associated genes, especially for some extreme circumstances such as polar night in circumpolar regions.

Sunshine duration is longer in subtropical latitudes (about 25° to 40° north/south) and shorter in latitudes > 50°, the annual sunshine hours difference can be up to 4-fold (Pinna M. 1978). Early modern humans firstly migrated out of Africa and steadily settled down across the whole world; thus, the originally adapted environment was abundant of sunshine (2400–4000 hour/year in Africa, Pinna M. 1978).

1.7 The purpose of this study

Human ancestors started to colonize Europe around 40 thousand years ago. At the end of Ice Age, which is 11,000 years ago, Mesolithic Pulli settlement located Estonia (Subrenat J. 2004). Not long after that, 9,500 years ago Mesolithic hunter-gatherer reached Scandinavia (Comparative timeline of y-DNA & mtDNA haplogroups development in Western Eurasia. Eupedia.com, Günther T et al. 2018). In Finland, pottery and agriculture were found of 7,300 years ago ((Humanistinen tiedekunta, Helsingin Yliopisto. Helsinki.fi. 2013). This timeline indicates north Europeans lived under a drastically less sunshine environment (1600–1800 hour/year or less, Pinna M. 1978) for thousands of years.

Knowing that the sunshine duration associated with myopia prevalence, have the human genome had been shaped by the ancestry environment variations through time? To verify the association, European is an ideal model for the diversity both in population and sunshine duration variations.

To find out whether living under different sunshine duration regions associated with myopia genetic risk, four European cohorts were selected due to their distinct annual sunshine duration differences (Pinna M. 1978). The myopia risk comparison will be conducted by PRS, the diversity of sunshine duration in Europeans and large-scale European centric GWASs make this study feasible.

2 Subjects and Methods

2.1 Subjects of this study

- Base data

Publicly available GWAS summary statistics containing 8,754,054 variants (Hysi PG et al. 2020) used as base data in this study. The original cohorts of the GWAS study including 276,065 European participants, which are two UK Biobank (UKBB) subjects (N=102,117 and 108,956 respectively), self-reported non-Hispanic white cohorts (N=34,998) from the Genetic Epidemiology Research on Adult Health and Aging (GERA), and non-UK participants of the Consortium for Refractive Error and Myopia (CREAM, N=29,994). The summary statistics data were downloaded from the FTP site of the King's College London (ftp://twinr-ftp.kcl.ac.uk/Refractive_Error_MetaAnalysis_2020).

Variants in the GERA and CREAM cohorts were imputed by using the cosmopolitan 1000 Genomes Project reference panel phase I integrated release (Hysi PG et al. 2020), and the two UKBB cohorts were genotyped using combined Haplotype Reference Consortium (HRC) and UK10K (a rare and low-frequency variants project in the UK population, Walter K et al. 2015) reference panel. The raw base data was not provided with SNP ID, while SNP ID is critical in PRS calculation. An open access data base UK GWAS round 2 variants panel was used as reference (www.nealelab.is/uk-biobank/). The UK GWAS round 2 variants panel genotypes were imputed from HRC, UK10K and 1000 Genomes reference panels, which is expected to achieve the most resemble SNP ID with GERA, CREAM and the two UKBB cohorts.

Base data SNP ID imputed by matching variant positions (chr:pos) between base data and UK GWAS round 2 variants panel using R (version 4.0.3), 7,582,481 of 8,754,054 variants successfully imputed, after filtering duplicated IDs, 7,538,107 variants were retained.

- Target data 1

Chip-based genotype data of the 1000 Genomes Project Phase 3 populations were

used as the Target data 1. VCF files of a total of 2,458,861 variants genotypes produced using Illumina Human Omni 2.5 platform were downloaded from the FTP sites of the International Genome Sample Resource (IGSR, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/). Five cohorts were selected for analysis, which were Finnish in Finland (FIN, N=100), British in England and Scotland (GBR, N=104), Iberian populations in Spain (IBS, N=150), Toscani in Italy (TSI, N=112) and CEPH (The Centre d'Etude du Polymorphism Humain) Utah residents with Northern and Western European ancestry (CEU, N=183).

The initial four cohorts have distinct annual sunshine duration difference, which were Helsinki in Finland 1,858 hours (Finnish Meteorological Institute, 2015), London in UK 1,633 hours (Met Office. 2014), Barcelona in Spain 2,591 hours (Guía resumida del clima en España 1981–2010. 2012) and Rome in Italy 2,473 hours (CLINO Averages Listed for the station Roma Ciampino. 2011). CEU is a mixed northern and western European ancestry cohort as control (CEPH website: <http://www.cephb.fr>, O'Brien E et al. 1994).

- Target data 2

Genotype data of 80,855,802 variants that were produced by whole-genome resequencing of the 1000 Genomes Project phase 3 populations was used as Target data 2. VCF files were downloaded from open access of IGSR (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>). Four cohorts were selected for analysis, which were Finnish in Finland (FIN, N=96), British in England and Scotland (GBR, N=91), Iberian populations in Spain (IBS, N=100), and Toscani in Italy (TSI, N=107). Target data 2 is considered to carry out replication for target data 1 at higher resolution.

2.2 Quality control (QC)

Filtering by R version 4.0.3 and PLINK version 1.9 (Chang CC et al. 2015), quality control procedures followed a PRS guide manual (Choi SW et al. 2020), key steps include:

- Standard GWAS QC

- Ambiguous SNPs
- Mismatching SNPs
- Duplicated SNPs
- Sex chromosomes
- Sample overlap
- Relatedness

For the base data, quality control criteria were used to remove variants with minor allele frequency (MAF) less than 0.01, duplicated variants, ambiguous SNPs, and mismatching SNPs, and finally 6,404,424 variants were retained.

Target data 1 was filtered by successful genotyping rate > 0.95 , sample missingness < 0.1 , Hardy-Weinberg Equilibrium $P > 1 \times 10^{-6}$, heterozygosity within 3 standard deviations of the mean, and MAF $> 1\%$.

Target data 2 was filtered by successful genotyping rate > 0.99 , sample missingness < 0.1 , Hardy-Weinberg Equilibrium $P > 1 \times 10^{-6}$, heterozygosity within 3 standard deviations of the mean, MAF $> 1\%$.

Table 1 shows the numbers of individuals and variants passed the 2-rounds filtering.

2.3 Myopia risk prediction

To assess the myopia risk for each cohort in the target data sets, Polygenic Risk Scores (PRS, Purcell SM et al. 2009) were calculated per individual. GWAS resulted in a set of markers (usually SNPs) associated to the trait of interest, effect sizes were estimated for each marker's association with the trait. The markers and corresponding effect sizes then used to assign PRS (Dudbridge F. 2013). The PRS is obtained by:

$$PRS_{PT,j} = \sum_{i=1}^m \beta_i G_{i,j}$$

where the PRS of individual j is equal to the weighted sum of effect allele at all m markers, and m is determined by P value threshold (PT). Score of the marker i equals to effect size β_i times the corresponding genotype score G_i , which given 0, 1 or 2 for times of effect allele appears at both chromosomes. In the standard PRS approach,

GWAS summary statistics will be applied with another set of training data to obtain the best-fit PT. Some set of m alleles have the top correlation (r^2) with the trait of interest, then calculate PRS at this PT for each individual in target data. In the meantime, the training data have to be independent from GWAS discovery samples to avoid overfitting (Choi SW et al. 2020).

However, PRS prediction accuracy decays with increasing genetic divergence between base and target data (Martin AR et al. 2017). Considering the base data is mainly originated from UK biobank British (76.5%), arbitrary PTs were used to avoid cohort bias in the standard approach (Choi SW et al. 2020). In the current study PRS were obtained at 8 PTs (Table 2) for target data 1 and 3 PTs (Table 2) for target data 2.

To further explore the data association, the relatively better PTs needed to be confirmed. Principal Component Analysis (PCA) was conducted using GCTA (Genome-wide Complex Trait Analysis, Yang et al. 2011) for all individuals in target data 1. The covariance of PRSs and 4 PCs at all 8 PTs in target data 1 results showing $PT=5 \times 10^{-3}$ as the best for the appropriate number of variants that keeps as many myopia-influenced variants and excludes as many non-influenced variants (Fig. 5, Fig. 1). This pattern successfully replicated in BMI using UK GWAS round 2 summary statistics (<http://www.nealelab.is/uk-biobank/>) and target data 1 as training data (Fig. 1).

2.4 Ancestry proportion of myopia associated variants

Aside from PRS as the aggregation of genetic effects, ancestry proportions of the associated variants could reveal the similarity among cohorts for myopia. Meanwhile, the trait specificity could be also verified.

Analysis carried by ADMIXTURE (Alexander DH et al. 2011), inputting base data and target data 1, at $PT=5 \times 10^{-3}$ extracting 10,253 scored variants. Outer group using YRI (Yoruba in Ibadan, Nigeria) from target data 1 raw panel (Fig. 6). Reference analysis conducted on target data 2 extracted 569,974 variants (Fig. 7).

3 Results

3.1 Polygenic risk scores

Summary statistics of a large-scale European centric myopia GWAS (N=276,065, Hysi PG et al. 2020) were used as base data to obtain PRSs in the current study. PRSs were firstly calculated per individual at 8 PTs (Table 2, Fig. 2) for target data 1, which consisted of the 1000 Genome Finish (FIN, N=98), British (GBR, N=94), Spanish (IBS, N=99), Italian (TSI, N=97) and Utah residents with Northern and Western European ancestry (CEU, N=101). Except for PT at 0.1, 0.05 and 0.01, PRSs in 5 of 8 PTs are following the similar pattern that IBS ranking the top and TSI ranking the bottom. Plus, PRSs refer to being less myopic and minus PRSs refer to being myopic.

PRSs calculation replication was carried out at 3 PTs (Table 2, Fig.3) for target data 2, which consisted of the 1000 Genome Finish (FIN, N=96), British (GBR, N=91), Spanish (IBS, N=100), Italian (TSI, N=107) with a 32-fold higher resolution than target data 1 (2.45 million and 80.85 million variants, respectively). Similar patterns replicated in the PRSs orders that IBS ranking top at PT of 5×10^{-3} and 5×10^{-8} , and TSI ranking the bottom at PT of 5×10^{-3} and 0.1.

These patterns are neither in accordance with the general genetic distance (Fig. 4) among cohorts nor the sunshine durations, which IBS and TSI are geographically closer (Pinna M. 1978). Notice that the relative genetic distance distribution, geographic distribution (Fig. 4) and sunshine duration (Šúri M et al. 2007) of the 4 European cohorts are highly correlated.

3.2 Best-fit P value threshold estimation

The PRSs patterns were not consensus at all PTs, in the standard PRS P value thresholding methods, the best-fit PT is obtained by finding the most relevant variant set to the trait of interest. To approximate the best-fit PRS, a regression analysis will be performed between PRSs calculated at a range of P -value thresholds and then select the PRS that explains the highest phenotypic variance (Choi SW et al. 2020). A third data

set is required for out-of-sample prediction in this approach, but the base data is mainly originated from British (76.5%), and target data are more evenly distributed among ethnicities, to avoid ethnicity bias, PRS and PCA covariance analysis was conducted instead of the standard approach.

Since not all the scored variants actually influence the trait under study, PT was used to shrink effect size of non-influenced variants to zero. The best-fit PT is at when shrinks as many non-influenced variants as possible while keeping as many influenced variants. The PRS and PCA covariance was larger when PRS distribution is closer to general genetic distance (PC1, Fig. 1) at larger PTs, 0.1e.g., for more variants were included. As PT shrinks, the proportion of truly influenced variants was getting higher because influenced variants often have smaller P values, while as PT shrinks to be smaller than the true best-fit PT, the number of the variants was so small that PRS distribution randomly shifts by PTs (Fig. 5, Fig. 1), thus the best-fit PT was around where PRS and PCA covariance firstly overlap among different PCs during shrinking the PT. In this case, 5×10^{-3} was relatively the best-fit PT in the results (Fig 1A).

3.3 Ancestry proportion of myopia associated variants

To confirm the results, ancestry proportion analyses were carried out using ADMIXTURE (Alexander DH et al. 2011). For target data 1, there were 10,253 scored variants at $PT=5 \times 10^{-3}$, and distinct ethnic specificity was observed, while IBS and TSI should have been more similar (Fig. 6).

As a reference, all 569,974 scored variants of the four cohorts in target data 2 resulted in a more mottled pattern (Fig. 7); nevertheless, at $K=3$ the ancestry proportion distribution was similar to the population stratification (Fig. 4) as expected. This mottled pattern might be due to abundant rare alleles in target data 2 than target data 1 (Fig. 8).

3.4 Comparison of scored variants

In the PRS formula (Dudbridge F. 2013), effect size and genotype of variants are

key variables. At population level, PRS is essentially determined by effect size and allele frequency. Since PRSs order were inconsistent with general genetic distance among the four cohorts, especially the genetic distance between IBS and TSI is the smallest (Table 3) but the PRS of them ranking top and bottom respectively. Venn Diagram Analysis (Fig. 9A) was carried out by R on target data 1 to identify the variants that played the major role. Target data 2 contained more rare variants (MAF < 0.1, Fig. 8), while the differences of these rare variants among cohorts are also small.

At P value < 5×10^{-3} , 30,765 variants of the four cohorts were taken into PRS scoring, among which 5,696 variants are mutual, and 2,243 variants are non-overlapped (Fig. 9A). For the 5,696 mutual variants, the Fixation indexes (Holsinger et al. 2009) were ranging from 0.0032 to 0.0076 (Table 3), even less than the general distance among the 4 cohorts (Nelis M et al. 2009), which could be considered as insignificant differentiation.

PRS formula is calculated for individual, as for cohort comparison, the formula can be transformed by:

$$\begin{aligned}
 PRS_{PT,j} &= \sum_{i=1}^m \beta_i G_{i,j} \\
 freq_i &= \frac{n_i}{2N} \\
 mean\ PRS_{PT,N} &= \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^m \beta_i G_{i,j} = \sum_{i=1}^m \beta_i \left(\frac{1}{N} \sum_{j=1}^N G_{i,j} \right) = \frac{1}{2} \sum_{i=1}^m \beta_i freq_i
 \end{aligned}$$

And in target data, the effect allele was the minor allele by default, thus by comparing the aggregation of $\beta \times \text{MAF}$ for all scored variants by cohort, which variants impacted the PRS order can be revealed. Results showing IBS is higher in partially and non-partially overlapped variants, and IBS is the only cohort $\sum \beta \times \text{MAF} > 0$ in non-overlapped variants (Fig. 9B). It was implying within the non-overlapped and partially overlapped variants; IBS has a more positive effect as a whole. This variant heterogeneity may potentially manipulate the PRS order.

3.5 Identifying the unique variants

Refer to non-overlapped variants (Fig. 9A), some of them are actual unique variants, but others are byproduct of linkage disequilibrium (LD) filtering during quality control and PRS calculation. The unique variant MAF must be no more than 0.01 in at least one cohort so that it was able to be filtered out during the PRS calculation. To identify them, the 2,243 SNP list were extracted in raw target data 1, and within each cohort, the numbers of variants $MAF \leq 0.01$ were 18 in FIN, 8 in GBR, 8 in IBS and 9 in TSI. After excluding duplicated ones, retained 32 true unique variants (Table 5), while some MAFs were extremely small, and the variations among cohorts were also negligible. Only 3 variants that have at least one cohort MAF larger than 10% (Table 5), rs1053046 ($P = 2.42 \times 10^{-3}$) in chromosome 6, a *PPARD* (peroxisome proliferator activated receptor delta) 3 Prime UTR variant; rs2844567 ($P = 2.81 \times 10^{-4}$) in chromosome 6 and rs17086731 ($P = 2.17 \times 10^{-3}$) in chromosome 13 have no specific genetic consequence (NIH dbSNP, www.ncbi.nlm.nih.gov/snp/).

The *PPARD* 3 Prime UTR variant rs1053046 frequency is 49.85% in African (Table 5) and steadily shrinks to Middle-East (Qatari 16.7%), southern European (IBS 11.33%, TSI 7.8%), and becomes extremely rare in northern European (~1%) presenting a geographical correlation. Knockout studies of *PPARD* in mice suggested a role for this protein in myelination of the corpus callosum, lipid metabolism, and epidermal cell proliferation (NCBI, www.ncbi.nlm.nih.gov/gene/5467). There is no evidence that rs1053046 is correlated with previously identified myopia-associated light signal and central nerve pathways (Hysi PG et al. 2020)

4 Discussion

4.1 The heterogeneity of myopia associated-variants

For monogenic traits, the ethnic specificities could be intuitively clear, such as aldehyde dehydrogenase 2 (*ALDH2*), the *ALDH2*2* (rs671) gene variant is absent among Europeans but is prevalent in populations in East Asian (Zhong Z et al. 2018). While for polygenic traits such as myopia, only a handful of loci were reported to show ethnic specificity (Yoshikawa M et al. 2014), the current study revealed myopia-associated variants vary across populations within Europeans. The population heterogeneity within Europeans is not negligible. For all scored variants of target data 1 (Table 1), the mutual variants proportion is no more than 40% (35,165 mutual variants). Due to the LD-based filtering procedure in PRS calculation, some non-overlapped variants were potential mutual ones (non-overlapped variants ratio, FIN 21.87%, GBR 17.47%, IBS 19.29% and TSI 19.63%), makes the maximum mutual ratio no more than 60% among cohorts.

The genetic differentiation of myopia-associated variants within Europeans was highly geographical. The 3 unique variants (rs2844567, rs1053046 and rs17086731) MAFs in different populations provide evidence for this correlation (NIH dbSNP, www.ncbi.nlm.nih.gov/snp/, Table 5). MAFs of all 3 variants were in perfect order rising from Northern European to Southern European, which are adjacent to North Africa or Middle-east whose MAFs are even higher, the gene flow across these continents has been reported (Hernández CL et al. 2020).

Except for environmental factors, historical factors also played an important role in genetic heterogeneity. Among the four cohorts, FIN was supposed to be the most peculiar cohorts for its relatively further genetic distance, while IBS stands out in the results. The genetic impacts of population movements left traces in Iberians, which were proposed to be associated with the Muslim conquest and the subsequent Reconquista (Bycroft C et al. 2019). The regionally varying fractions of north-west African ancestry (0–11%) in modern-day Iberians were reported, related to an

admixture event involving European-like and north-west African-like source populations, and these impacts may induce a part of myopia-associated variants variation.

4.2 Selective pressure and sunshine duration hypothesis

Natural selection can purge deleterious genotypes and thus alleles with large effects on a phenotype are found to be rare in a population. This pattern was observed by comparing the effect size and frequency distribution of variants in several polygenic traits (Fig. 10). The significance of variants is often correlated with the effect size; for traits under strong negative selective pressure, variants with large effect sizes often showed low MAF. For instance, Type-1 Diabetes and schizophrenia showed the excess of rare variants with large effects, suggesting the presence of strong selective pressure. On the other hand, left-handedness and height showed a moderate distribution. Myopia and body mass index (BMI) shared normal-distribution-like patterns that stand for weak or absence of selective pressure.

Previous GWASs have identified all anatomical components of the eye and central nervous system that participated in the myopia development, most of the associated loci are within intron or intergenic regions, and not likely to be subject to particularly strong selective pressures (Hysi PG et al. 2020). The normally distributed like variants frequencies also supported this idea (Fig. 10). There was also hypothesis about the myopia genetic diversity, that it is the result of overall balancing forces that encourage high allelic diversity of genes providing additional buffering capacity to absorb environmental pressures (Hysi PG et al. 2020). This hypothesis could be one explain for the results in the current study, the variants vary while the aggregate effects (PRSs) differences are small.

The evolution of eye had been the major challenge of Darwin and eventually it has been proven that the functional structures and mechanism of eye and vision were developed through adaptation (Schwab IR. 2018). Myopia is essentially non-adapted to less and weak lighting. Since the Northern European had living under less sunshine for

thousands of years, there could be signature of adaptation to weak-light environment in their genomes, but critical evidence still needed. The myopia epidemic swept across Northern Europe including Finland and Sweden (Bourne RR et al. 2016, Morgan IG et al. 2012), while a Norwegian Adolescent myopia study found this trend absent even considered the education pressure is as strong as the rest of Europe (Hagen LA et al. 2018). Their study based on proposing in regions with less sunshine duration, people should have been more myopic, and trying to explain their observation by answering why daylight exposure during a relatively short summer outweighs that of the longer autumn-winter. This study could be a breakthrough in verifying whether adaptation has driven this case. Norwegian genetic data (Mattingsdal M et al. 2020) are expected to enrich the Northern European target data and make more precise Northern European comparison and specificity identification possible.

4.3 Indication from the results

The results were not able to confirm the association between ancestry sunshine duration and myopia risk. Still, on the other hand, they revealed the complexity of genome diversity involved in risks of myopia. Although it is difficult to develop effective treatment base on genetic methods, given the strong impact of environmental factors and small aggregate genetic risk differences (PRSs) among populations, the evolutionary mismatch of myopia is expected to be relieved.

Meanwhile, whether similar heterogeneity of myopia associated variants patterns exist in other polygenic diseases remains unknown. Thus, it needs to be cautious with population specificity when developing targeted therapies for polygenic traits, the random variations could exist within the same ethnicity. The potential heterogeneity of common polygenic diseases can be verified by the similar method in the current study, while the data with enough power and depth are critical, and the disparity of ethnicity remains an obstacle.

4.4 Challenge of this study

Except for higher myopia prevalence reported in East Asians (Holden BA et al. 2016), there were no other appreciable ethnicity features in the myopia epidemic. There is also little previous data and information about the relationship between latitude and myopia risk in various populations. The current study aimed to find an association between latitude, which is equivalent to sunshine duration, and genetic risk, but results were more complicated than expected.

Based on what've been learnt, the impact of environmental factors was strong and contribution of genomes to the ethnic differences in the myopia prevalence might not be clearly manifested. However, it is reasonable to presume that in myopia there could be ethnic differences in genetic risks that was shaped by local adaptation similar to other traits like skin colors.

The first challenge comes from base data, or discovery data, the heritability of myopia estimated to be around 0.2 (Tedja MS et al. 2018), that hundreds of thousands of participants are required to enhance the power and prediction accuracy while as the population size within a single ethnicity gets larger, there comes the problem of population stratification. The base data used in the current study was so far the largest one, sample size up to 542,934 while the genomic inflation factor $\lambda = 1.94$ which normally recommended to be under 1.3 (Hysi PG et al. 2020). The genomic inflation factor expresses the deviation of the distribution of the observed test statistic compared to the distribution of the expected test statistic. High genomic inflation factors are often caused by population stratification, strong linkage disequilibrium (LD) between SNPs, strong association between SNPs and phenotypes, and systematic bias (van den Berg S et al. 2019).

Secondly, the base data was from a mixed European cohort (76.5% British) GWAS, key variables such as allele frequencies were averaged among cohorts, thus the majority of ethnic-specific variants cannot preserve until PRS scoring, notice the GWAS only tells the mutual association to the trait of interest from discovery samples. Moreover, from the summary statistics, the effect sizes were fixed to the origin discovery samples,

but the target data do not necessarily share the same effect size at all scored variants. In the current study, the unexpected PRS order was observed, while the thorough differences among the four cohorts are so small that makes it difficult to compare these differences (Fig. 5).

Target data also limited prediction accuracy. Heritability (h^2) reflects the genetic contributions to a phenotypic variance and the h^2 of myopia is around 0.2 (Tedja MS et al. 2018), while the PRS manual recommends that a target sample size of ~200 for a trait of $h^2 = 0.23$ so the statistical power can exceed 80% (Choi SW et al. 2020). In the current study ~100 samples in each target data estimated to end up with no more than 60% statistical power according to the manual.

Nevertheless, the base data and target data were the latest compatible and accessible data, and European cohorts are ideal subjects for this study. The unique pattern of IBS does not necessarily indicate they will be less myopic under similar environmental exposure, but could be seen as another trace of ancestry.

In the following study, a Swedish genetic database (SweGen, Ameer A et al. 2017) will be reached out as a supplement, together with present data for more precise verification of latitude-myopia correlation analysis. Tohoku Medical Megabank Organization (ToMMO, www.megabank.tohoku.ac.jp) data will be considered for GWAS to obtain myopia summary statistics and combined with other East Asian cohorts to verify whether similar heterogeneity exists in Asian.

In summary, previous study proposed that myopia prevalence has a correlation with sunshine duration (Choo HG et al. 2019) and the current study tried to find the correlation between genetic risk of myopia and geography in Europe, which could be shaped by local adaptive evolution to different light environments. The current study applied myopia GWAS summary statistics to four European cohorts with different latitude for genetic risk comparison. PRSs indicated the genetic risk order was not in accordance with their general genetic distance and sunshine duration differences. By further comparing the PRS scored variants, the partially-overlapped and non-overlapped variants played the major role in PRS orders. This heterogeneity in myopia-

associated variants implies the complexity of polygenic disease and the genetic diversity within Europeans.

5 Acknowledgements

I would like to thank my supervisor professor Kazuhiro Nakayama for the advices and support throughout the current study. The idea of this study improvised from an inexperienced and bold conjuncture, the process of starting a project from ground zero is difficult, Professor Nakayama's encouragement and enlightenment helped me went through the most puzzled times, and it's of great enjoyment working with our lab mates. 2020 is a difficult time for people all over the world, as a member of Laboratory of Evolutionary Anthropology, Department of Integrated Bioscience, Graduate School of Frontier Science, The University of Tokyo, I'd like to express gratitude for all the care and help from members of the laboratory during my quarantine in China. This study is but a preceding part of attempts to answer a series of questions in the field, at the new dawn of the human genome era, I believe more exciting achievements are in prospect with our Laboratory.

6 References

- Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12: 246
- Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kähäri AK, Lundin P, Che H, Thutkawkorapin J, Einfeldt J, Lampa S, Dahlberg M, Hagberg J, Jareborg N, Liljedahl U, Jonasson I, Johansson Å, Feuk L, Lundeberg J, Syvänen AC, Lundin S, Nilsson D, Nystedt B, Magnusson PK, Gyllensten U. 2017. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur J Hum Genet.* 25(11): 1253-1260
- Ashby RS, Schaeffel F. 2010. The effect of bright light on lens compensation in chicks. *Invest Ophthalmol Vis Sci* 51(10): 5247-5253
- Bayoumi R, De Fanti S, Sazzini M, Giuliani C, Quagliariello A, Bortolini E, Boattini A, Al-Habori M, Al-Zubairi AS, Rose JI, Romeo G, Al-Abri A, Luiselli D. 2016. Positive selection of lactase persistence among people of Southern Arabia. *Am J Phys Anthropol* 161(4): 676-684
- Bourne RR, Stevens GA, White RA, Smith JL, Flaxman SR, Price H, Jonas JB, Keeffe J, Leasher J, Naidoo K, Pesudovs K, Resnikoff S, Taylor HR. 2013. Vision Loss Expert Group. Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health* 1(6): e339-349
- Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, Quintela I, Carracedo A, Donnelly P, Myers S. 2019. Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun* 551
- Chan R, Trobe J. 2002. Spasm of accommodation associated with closed head trauma. *J Neuroophthalmol.* 22 (1): 15–17
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4: 7
- Chial H. 2008. Rare Genetic Disorders: Learning About Genetic Disease Through Gene Mapping, SNPs, and Microarray Data. *Nature Education* 1(1): 192

- Choi, SW, Mak TS & O'Reilly PF. 2020. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 15: 2759-2772
- Choo HG, Rah SH, Kim SH. 2019. Comparison of Solar Radiation and Myopia Occurrence in South Korean Children. *Journal of Ophthalmology*. Artrical ID 7643850
- Cline, D., Hofstetter, H. W., & Griffin, JR. 1997. *Dictionary of Visual Science* (4th ed). Butterworth-Heinemann, Boston.
- Croke K, Ishengoma DS, Francis F, Makani J, Kamugisha ML, Lusingu J, Lemnge M, Larreguy H, Fink G, Mmbando BP. 2017. Relationships between sickle cell trait, malaria, and educational outcomes in Tanzania. *BMC Infect Dis*. 17(1):568
- Deng L, Xu S. 2017. Adaptation of human skin color in various populations. *Hereditas* 155, 1
- Dolgin E. 2015. The myopia boom. *Nature* 519:276-278
- Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet*. 9(3): e1003348
- Foster PJ, Jiang Y. 2014. Epidemiology of myopia. *Eye (Lond)* 28(2): 202-208
- Glazier AM, Nadeau JH, Aitman TJ. 2002. Finding genes that underlie complex traits. *Science*. 298(5602): 2345-9.
- DA Goss, Grosvenor TP, Keller JT, Marsh-Tootle W, Norton TT, Zadnik K. 1997. *Optometric Clinical Practice Guideline: Care of the Patient with Myopia*. American Optometric Association 6
- Grosvenor T. 1987. A review and a suggested classification system for myopia on the basis of age-related prevalence and age of onset. *Optometry and Vision Science* 545–554
- Günther T, Malmström H, Svensson EM, Omrak A, Sánchez-Quinto F, Kılınç, Gülşah M, Krzewińska M, Eriksson G, Fraser M. 2018. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLOS Biology* 16(1): e2003703
- Hagen LA, Gjelle JVB, Arnegard S, Pedersen HR, Gilson SJ, Baraas RC. 2018.

- Prevalence and Possible Factors of Myopia in Norwegian Adolescents. *Sci Rep*. 8(1): 13479
- Hernández CL, Pita G, Cavadas B, López S, Sánchez-Martínez LJ, Dugoujon J, Novelletto A, Cuesta P, Pereira L, Calderón R. 2020. Human Genomic Diversity Where the Mediterranean Joins the Atlantic. *Molecular Biology and Evolution* 4: 1041–1055
- Holden BA, Fricke TR, Wilson DA, Jong M, Naidoo KS, Sankaridurg P, Wong TY, Naduvilath TJ, Resnikoff S. 2016. Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050. *Ophthalmology* 123(5): 1036-1042
- Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nat Rev Genet*. 10(9): 639-650
- Hysi PG, Choquet H, Khawaja AP, Wojciechowski R, Tedja MS, Yin J, Simcoe MJ, Patasova K, Mahroo OA, Thai KK, Cumberland PM, Melles RB, Verhoeven VJM, Vitart V, Segre A, Stone RA, Wareham N, Hewitt AW, Mackey DA, Klaver CCW, MacGregor S; Consortium for Refractive Error and Myopia, Khaw PT, Foster PJ; UK Eye and Vision Consortium, Guggenheim JA; 23andMe Inc., Rahi JS, Jorgenson E, Hammond CJ. 2020. Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia. *Nat Genet* 52(4): 401-407
- Hysi PG, Wojciechowski R, Rahi JS, Hammond CJ. 2014. Genome-wide association studies of refractive error and myopia, lessons learned, and implications for the future. *Invest Ophthalmol Vis Sci* 55: 3344–3351.
- Knowles JW, Ashley EA. 2018. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med* 15(3): e1002546
- Lagrèze WA, Schaeffel F. 2017. Preventing Myopia. *Dtsch Arztebl Int* 114: 575-580
- Landis EG, Yang V, Brown DM, Pardue MT, Read SA. 2018. Dim Light Exposure and Myopia in Children. *Investigative Ophthalmology & Visual Science* 59: 4804-4811

- Leslie R, O'Donnell CJ, Johnson AD. 2014. GRASP: analysis of genotype-phenotype results from 1,390 genome-wide association studies and corresponding open access database. *Bioinformatics* 30(12): i185-94
- Long E. 2018. Evolutionary medicine Why does prevalence of myopia significantly increase? *Evol Med Public Health* 1: 151–152.
- Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 51(4):584-591
- Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* 100(4): 635-649
- Mattingdal M, Ebenesersdóttir SS, Moore KSH, Andreassen OA, Hansen TF, Werge T, Kockum, Olsson T, Alfredsson L, Helgason H, Stefánsson K, Hovig E. 2020. *bioRxiv* 2020.03.20.000299
- Morgan I, Rose K. 2005. How genetic is school myopia? *Progress in Retinal and Eye Research* 24(1): 1-38
- Morgan IG, French AN, Ashby RS, Guo X, Ding X, He M, Rose KA. 2018. The epidemics of myopia: Aetiology and prevention. *Progress in Retinal and Eye Research* 62: 134–149.
- Morgan IG, Rose KA. 2018. Myopia: is the nature-nurture debate finally over? *Clinical and Experimental Optometry*. 102(1): 3-17
- Morgan IG, Ohno-Matsui K, Saw SM. 2012. Myopia. *Ophtalmology The Lancet* 379: 1739-1748
- Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskácková T, Balascák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A,

- Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. 2009. Genetic structure of Europeans: a view from the North-East. *PLoS One* 4(5): e5472
- O'Brien E, Rogers AR, Beesley J, Jorde LB. 1994. Genetic structure of the Utah Mormons: Comparison of results based on RFLPs, blood groups, migration matrices, isonymy, and pedigrees. *Human Biology* 66(5): 743-759
- Pinna M. 1978. *L'atmosfera e il clima*. Torino: UTET 478
- Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460(7256): 748-752
- Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Grønnow B, Appelt M, Gulløv HC, Friesen TM, Fitzhugh W, Malmström H, Rasmussen S, Olsen J, Melchior L, Fuller BT, Fahrni SM, Stafford T Jr, Grimes V, Renouf MA, Cybulski J, Lynnerup N, Lahr MM, Britton K, Knecht R, Arneborg J, Metspalu M, Cornejo OE, Malaspina AS, Wang Y, Rasmussen M, Raghavan V, Hansen TV, Khusnutdinova E, Pierre T, Dneprovsky K, Andreassen C, Lange H, Hayes MG, Coltrain J, Spitsyn VA, Götherström A, Orlando L, Kivisild T, Villems R, Crawford MH, Nielsen FC, Dissing J, Heinemeier J, Meldgaard M, Bustamante C, O'Rourke DH, Jakobsson M, Gilbert MT, Nielsen R, Willerslev E. 2014. The genetic prehistory of the New World Arctic. *Science* 345(6200): 1255832
- Saw SM, Gazzard G, Shih-Yen EC, Chua WH. 2005. Myopia and associated pathological complications. *Ophthalmic Physiol* 25(5): 381-91.
- Schwab IR. 2018. The evolution of eyes: major steps. The Keeler lecture 2017: centenary of Keeler Ltd. *Eye (Lond)*. 32(2): 302-313
- Smith TST, Frick KD, Holden BA, Fricke TR, Naidoo KS. 2009. Potential lost productivity resulting from the global burden of uncorrected refractive error. *Bull World Health Org* 87: 431-437

- Šúri M, Huld TA, Dunlop ED, Ossenbrink HA. 2007. Potential of solar electricity generation in the European Union member states and candidate countries. *Solar Energy*, 81: 1295–1305
- Tedja MS, Wojciechowski R, Hysi PG, Eriksson N, Furlotte NA, Verhoeven VJM, Iglesias AI, Meester-Smoor MA, Thompson SW, Fan Q, Khawaja AP, Cheng CY, Höhn R, Yamashiro K, Wenocur A, Graza C, Haller T, Metspalu A, Wedenoja J, Jonas JB, Wang YX, Xie J, Mitchell P, Foster PJ, Klein BEK, Klein R, Paterson AD, Hosseini SM, Shah RL, Williams C, Teo YY, Tham YC, Gupta P, Zhao W, Shi Y, Saw WY, Tai ES, Sim XL, Huffman JE, Polašek O, Hayward C, Bencic G, Rudan I, Wilson JF; CREAM Consortium; 23andMe Research Team; UK Biobank Eye and Vision Consortium, Joshi PK, Tsujikawa A, Matsuda F, Whisenhunt KN, Zeller T, van der Spek PJ, Haak R, Meijers-Heijboer H, van Leeuwen EM, Iyengar SK, Lass JH, Hofman A, Rivadeneira F, Uitterlinden AG, Vingerling JR, Lehtimäki T, Raitakari OT, Biino G, Concas MP, Schwantes-An TH, Igo RP Jr, Cuellar-Partida G, Martin NG, Craig JE, Gharahkhani P, Williams KM, Nag A, Rahi JS, Cumberland PM, Delcourt C, Bellenguez C, Ried JS, Bergen AA, Meitinger T, Gieger C, Wong TY, Hewitt AW, Mackey DA, Simpson CL, Pfeiffer N, Pärssinen O, Baird PN, Vitart V, Amin N, van Duijn CM, Bailey-Wilson JE, Young TL, Saw SM, Stambolian D, MacGregor S, Guggenheim JA, Tung JY, Hammond CJ, Klaver CCW. 2018. Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet.* 50(6): 834-848
- van den Berg S, Vandenplas J, van Eeuwijk FA, Lopes MS, Veerkamp RF. 2019. Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data. *J Anim Breed Genet.* 136(6):418-429
- Verhoeven VJ, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, Höhn R, MacGregor S, Hewitt AW, Nag A, Cheng CY, Yonova-Doing E, Zhou X, Ikram MK, Buitendijk GH, McMahon G, Kemp JP, Pourcain BS, Simpson CL, Mäkelä KM, Lehtimäki T, Kähönen M, Paterson AD, Hosseini SM, Wong HS, Xu L,

- Jonas JB, Pärssinen O, Wedenoja J, Yip SP, Ho DW, Pang CP, Chen LJ, Burdon KP, Craig JE, Klein BE, Klein R, Haller T, Metspalu A, Khor CC, Tai ES, Aung T, Vithana E, Tay WT, Barathi VA; Consortium for Refractive Error and Myopia (CREAM), Chen P, Li R, Liao J, Zheng Y, Ong RT, Döring A; Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Research Group, Evans DM, Timpson NJ, Verkerk AJ, Meitinger T, Raitakari O, Hawthorne F, Spector TD, Karssen LC, Pirastu M, Murgia F, Ang W; Wellcome Trust Case Control Consortium 2 (WTCCC2), Mishra A, Montgomery GW, Pennell CE, Cumberland PM, Cotlarciuc I, Mitchell P, Wang JJ, Schache M, Janmahasatian S, Igo RP Jr, Lass JH, Chew E, Iyengar SK; Fuchs' Genetics Multi-Center Study Group, Gorgels TG, Rudan I, Hayward C, Wright AF, Polasek O, Vataavuk Z, Wilson JF, Fleck B, Zeller T, Mirshahi A, Müller C, Uitterlinden AG, Rivadeneira F, Vingerling JR, Hofman A, Oostra BA, Amin N, Bergen AA, Teo YY, Rahi JS, Vitart V, Williams C, Baird PN, Wong TY, Oexle K, Pfeiffer N, Mackey DA, Young TL, van Duijn CM, Saw SM, Bailey-Wilson JE, Stambolian D, Klaver CC, Hammond CJ. 2013. Genome-wide meta-analyses of multi-ancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet.* 45(3): 314-8
- Vitale S, Robert DS, Ferris FL III. 2009. Increased prevalence of myopia in the United States between 1971-1972 and 1999-2004. *Arch Ophthalmol* 127(12): 1632-1639
- Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Richards JB, Greenwood CM, Timpson NJ, Durbin R, Soranzo N. UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* 526: 82-90
- Williams KM, Verhoeven VJ, Cumberland P, Bertelsen G, Wolfram C, Buitendijk GH,

- Hofman A, van Duijn CM, Vingerling JR, Kuijpers RW, Höhn R, Mirshahi A, Khawaja AP, Luben RN, Erke MG, von Hanno T, Mahroo O, Hogg R, Gieger C, Cougnard-Grégoire A, Anastasopoulos E, Bron A, Dartigues JF, Korobelnik JF, Creuzot-Garcher C, Topouzis F, Delcourt C, Rahi J, Meitinger T, Fletcher A, Foster PJ, Pfeiffer N, Klaver CC, Hammond CJ. 2015. Prevalence of refractive error in Europe: the European Eye Epidemiology (E(3)) Consortium. *Eur J Epidemiol.* 30(4): 305-315
- Williams KM, Bertelsen G, Cumberland P, Wolfram C, Verhoeven VJ, Anastasopoulos E, Buitendijk GH, Cougnard-Grégoire A, Creuzot-Garcher C, Erke MG, Hogg R, Höhn R, Hysi P, Khawaja AP, Korobelnik JF, Ried J, Vingerling JR, Bron A, Dartigues JF, Fletcher A, Hofman A, Kuijpers RW, Luben RN, Oxele K, Topouzis F, von Hanno T, Mirshahi A, Foster PJ, van Duijn CM, Pfeiffer N, Delcourt C, Klaver CC, Rahi J, Hammond CJ; European Eye Epidemiology (E(3)) Consortium. 2015. Increasing Prevalence of Myopia in Europe and the Impact of Education. *Ophthalmology* 122(7): 1489-97
- Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. 2013. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet.* 14(7): 507-15.
- Xiong S, Sankaridurg P, Naduvilath T, Zang JJ, Zou HD, Zhu JF, Lv MZ, He XG, Xu X. 2017. Time spent in outdoor activities in relation to myopia prevention and control: a meta-analysis and systematic review. *Acta Ophthalmol* 95(6): 551-566
- Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1): 76-82.
- Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM; GIANT Consortium. 2018. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet.* 27(20): 3641-3649
- Yoshikawa M, Yamashiro K, Miyake M, Oishi M, Akagi-Kurashige Y, Kumagai K, Nakata I, Nakanishi H, Oishi A, Gotoh N, Yamada R, Matsuda F, Yoshimura N;

- Nagahama Study Group. 2014. Comprehensive replication of the relationship between myopia-related genes and refractive errors in a large Japanese cohort. *Invest Ophthalmol Vis Sci.* 55(11): 7343-7354
- Zhong Z, Hou J, Li B, Zhang Q, Li C, Liu Z, Yang M, Zhong W, Zhao P. 2018. Genetic Polymorphisms of the Mitochondrial Aldehyde Dehydrogenase ALDH2 Gene in a Large Ethnic Hakka Population in Southern China. *Med Sci Monit.* 24: 2038-2044
- Zhou Z, Chen T, Wang M, Jin L, Zhao Y, Chen S, Wang C, Zhang G, Wang Q, Deng Q, Liu Y, Morgan IG, He M, Liu Y, Congdon N. 2017. Pilot study of a novel classroom designed to prevent myopia by increasing children's exposure to outdoor light. *PLoS ONE* 12(7): e0181772

Table 1. Individual and variant numbers of target data 1 and 2 after 2 rounds QC.

Target data 1	N	ALL variants taken into score after 2 rounds filtering
FIN	98	83,299
GBR	94	91,544
IBS	99	96,202
TSI	97	97,033
CEU	101	92,426
raw		2,458,861
Target data 2	N	ALL variants taken into score after 2 rounds filtering
FIN	96	216,361
GBR	91	245,341
IBS	100	262,336
TSI	107	251,700
raw		80,855,702

Table 2 Arbitrary P-value threshold (PT) for target data PRS calculation.

Set number for target data 1	PT	Base data variant number
S1	5.00E-20	4,494
S2	5.00E-12	15,796
S3	5.00E-08	38,598
S4	5.00E-05	108,396
S5	5.00E-03	385,726
S6	0.01	505,902
S7	0.05	1,082,909
S8	0.1	1,588,735
raw		8,754,054
Set number for target data 2	PT	Base data variant number
S3	5.00E-08	38,598
S5	5.00E-03	385,726
S8	0.1	1,588,735
raw		8,754,054

Table 3 Fixation index between 4 European cohorts by 5,696 mutual variants.

Fst	GBR	IBS	TSI
FIN	0.0054±0.0076	0.0069±0.0098	0.0076±0.0107
GBR		0.0036±0.0053	0.0043±0.0060
IBS			0.0032±0.0046

Table 4 Unique variants MAF in scored variants at $PT=5 \times 10^{-3}$.

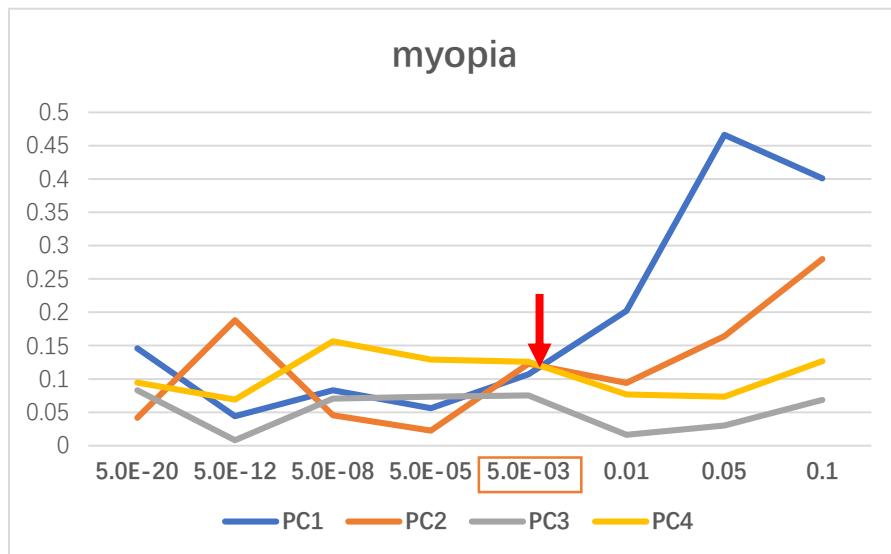
Ref.: reference allele (minor allele), Alt.: alternative allele.

SNP	CHR	Ref.	Alt.	FIN	GBR	IBS	TSI
rs1611772	1	A	G	4.50%	0.96%	4.36%	0.45%
rs543314	1	C	T	3.00%	1.44%	0.67%	3.60%
rs7594497	2	C	T	1.00%	4.33%	3.00%	3.64%
rs12611664	2	C	T	6.00%	0.48%	0.00%	0.45%
rs17727965	2	T	C	9.00%	3.85%	0.33%	2.23%
rs11689345	2	A	G	2.04%	2.91%	0.00%	0.45%
rs1001296	2	C	A	6.50%	3.85%	3.00%	0.89%
rs2573226	2	A	G	3.50%	2.89%	2.00%	0.45%
rs10007113	4	A	G	0.50%	2.40%	4.00%	4.46%
rs7687899	4	G	T	0.50%	1.92%	2.00%	3.18%
rs17355550	4	C	T	8.50%	3.37%	1.67%	0.89%
rs1150733	6	A	G	1.00%	5.29%	8.00%	1.34%
rs2844567	6	A	G	1.00%	8.65%	22.00%	8.93%
rs7743807	6	T	C	0.50%	6.25%	6.67%	2.23%
rs106287	6	A	G	0.50%	4.33%	5.03%	1.34%
rs204889	6	A	G	0.50%	4.41%	6.33%	1.80%
rs413887	6	C	A	0.50%	2.89%	8.00%	4.17%
rs1053046	6	A	G	1.00%	0.48%	11.33%	7.80%
rs7742752	6	A	G	1.52%	0.96%	3.36%	4.02%
rs6915101	6	T	C	1.50%	0.96%	3.69%	4.02%
rs11966753	6	G	T	5.00%	4.37%	1.00%	1.35%
rs17158486	7	T	C	0.00%	0.96%	0.33%	2.68%
rs11773581	7	A	G	5.50%	8.33%	4.00%	0.90%
rs4263805	8	T	C	0.00%	3.85%	5.33%	4.46%
rs7121134	11	T	G	2.00%	0.00%	3.00%	0.89%
rs17086731	13	A	G	1.00%	7.21%	9.06%	14.29%
rs17079019	13	A	G	0.50%	2.40%	5.33%	2.23%
rs2243110	17	A	G	0.50%	2.89%	0.67%	0.45%
rs1992507	17	T	C	1.00%	1.92%	1.00%	2.23%
rs9957716	18	C	T	0.00%	2.50%	2.63%	2.94%
rs16975268	18	C	T	1.00%	1.44%	4.00%	3.57%
rs2307277	19	G	A	2.50%	0.48%	1.68%	1.34%

Table 5 The 3 unique variants MAF compare to other adjacent populations.
MAF information from NIH dbSNP website, www.ncbi.nlm.nih.gov/snp.
Latitude and longitude from Maps of the world website, www.mapsofworld.com.
The MAF distribution shows a geographical predisposition.

	rs2844567	rs1053046	rs17086731	Latitude and longitude
Northern Europe				
FIN	1.00%	1.00%	1.00%	60°~70° N 25°~30° E
Estonian	3.06%	1.85%	4.02%	58°~60° N 25° E
Northern Sweden	2.70%	1.30%	6.20%	60°~70° N 15°~20° E
Western Europe				
GBR	8.65%	0.48%	7.21%	51°~59° N 0°~10° W
UK10K twins	7.96%	4.94%	7.28%	51°~59° N 0°~6° W
Southern Europe				
IBS	22.00%	11.33%	9.06%	36°~44° N 0°~8° W
TSI	8.93%	7.80%	14.29%	37°~46° N 7°~18° E
Middle East				
Qatari	22.20%	16.70%	9.70%	25°~26° N 51°~52° E
1000 genomes African	17.17%	49.85%	11.80%	20° N ~ 20° S 20° W ~ 40° E
1000 genomes European	10.14%	5.07%	8.05%	-

A.



B.

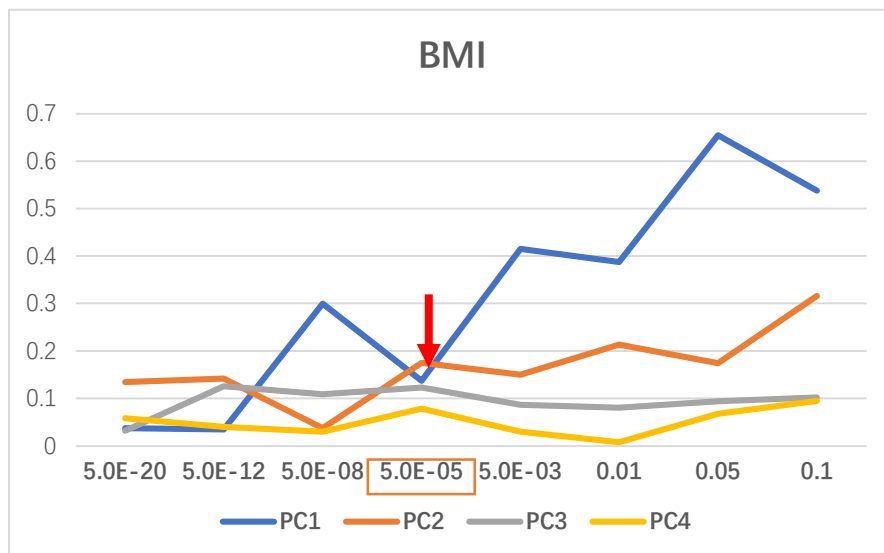
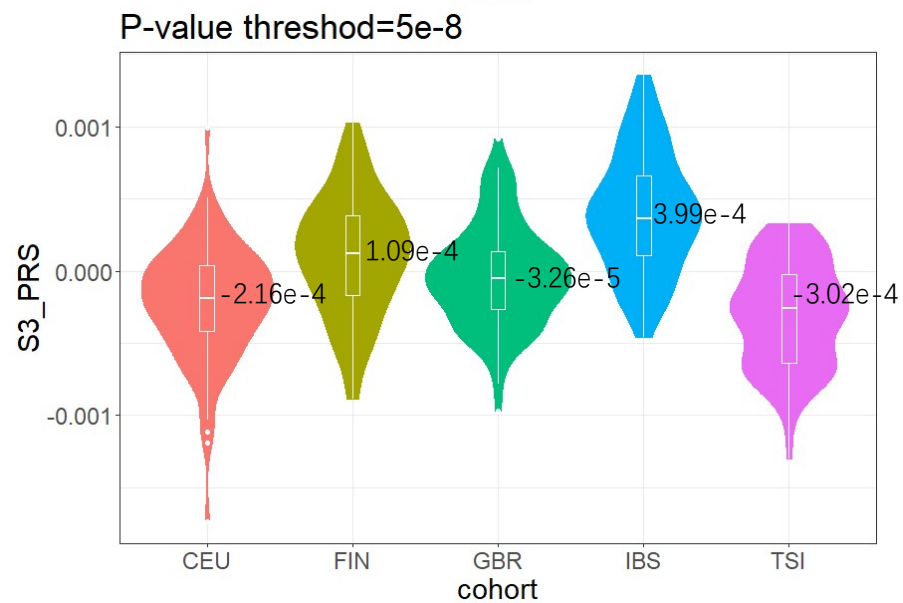
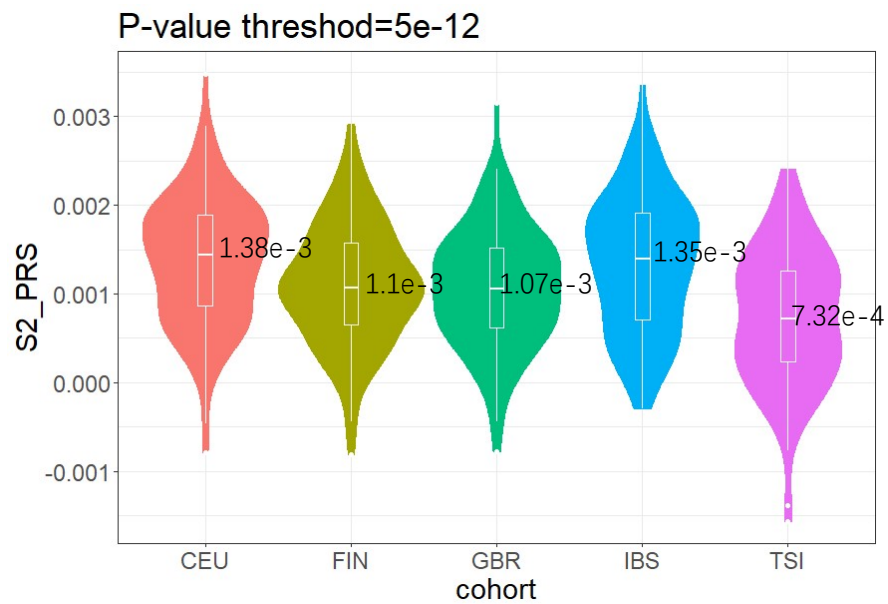
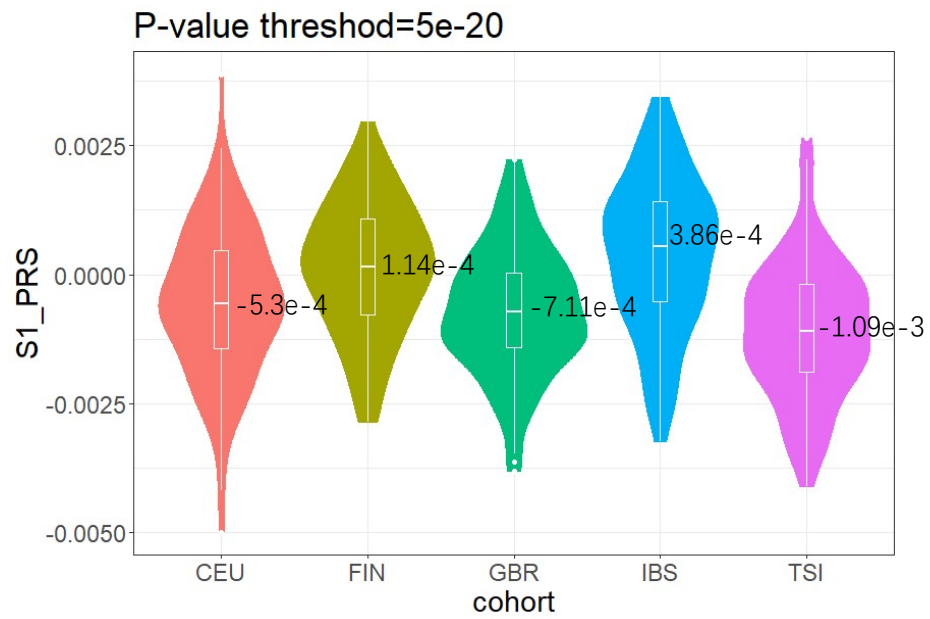
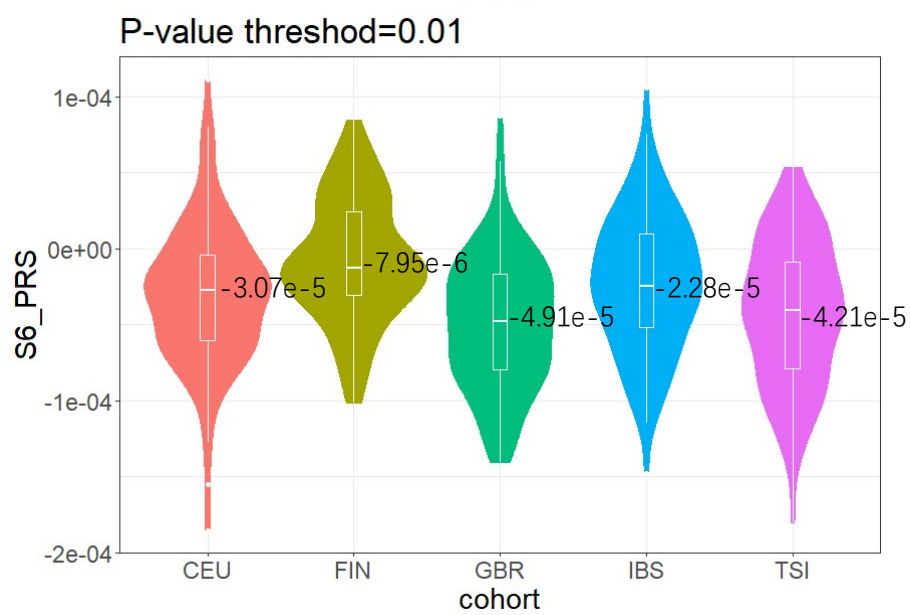
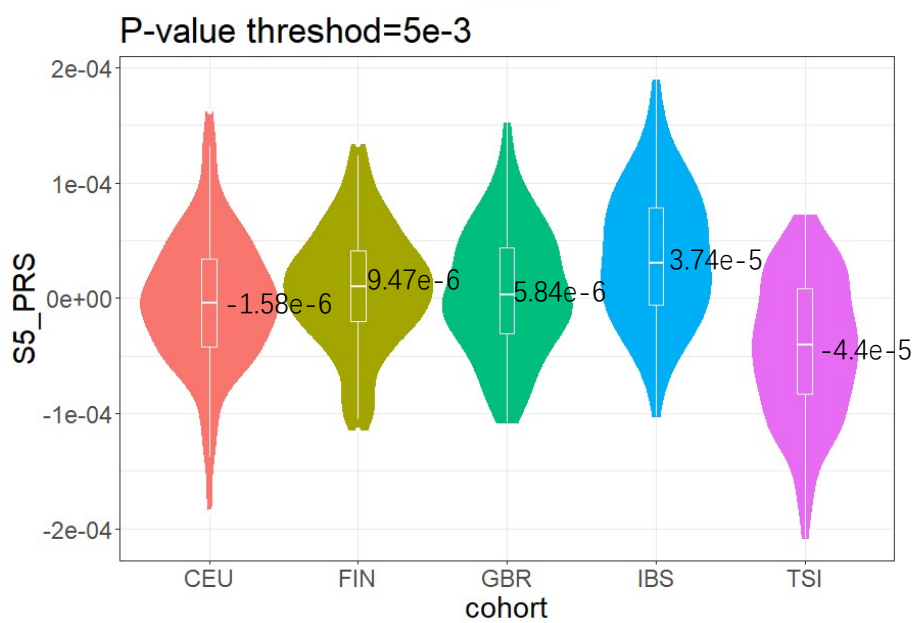
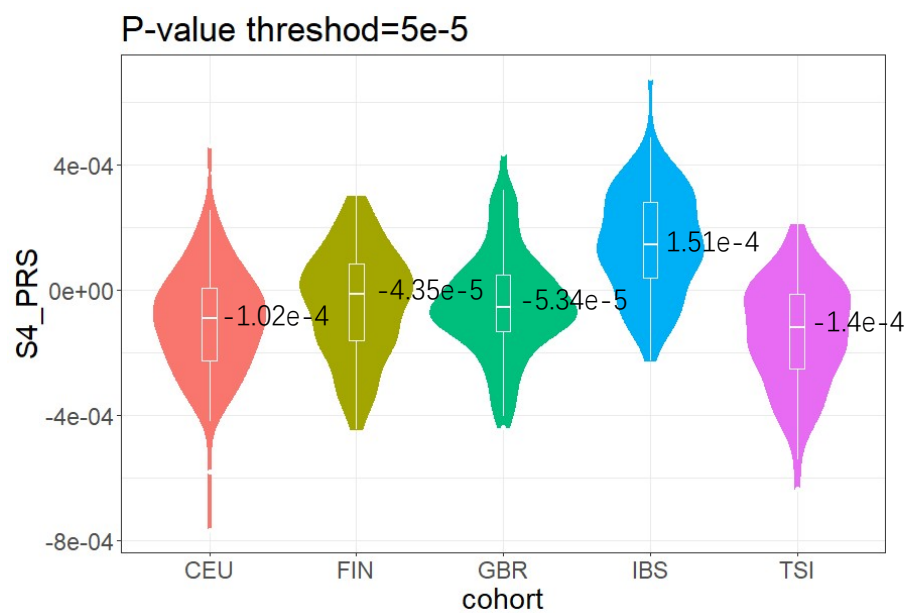


Fig. 1 PCA and PRSs results covariance distribution

Principle components and PRSs covariance distribution for myopia and BMI. A: PCs are obtained from all PRS scored variants in target data 1 individuals, at P value threshold 5×10^{-3} covariance crossed among PCs, PT determines the variants number for PRSs, the best PT is at maintaining in accordance with main PC (PC1) while excluding as many possible non-influenced variants. B: replication for BMI PRSs and PCA covariance distribution, individual data and summary statistics were from UK Biobank GWAS round 2 (<http://www.nealelab.is/uk-biobank/>).





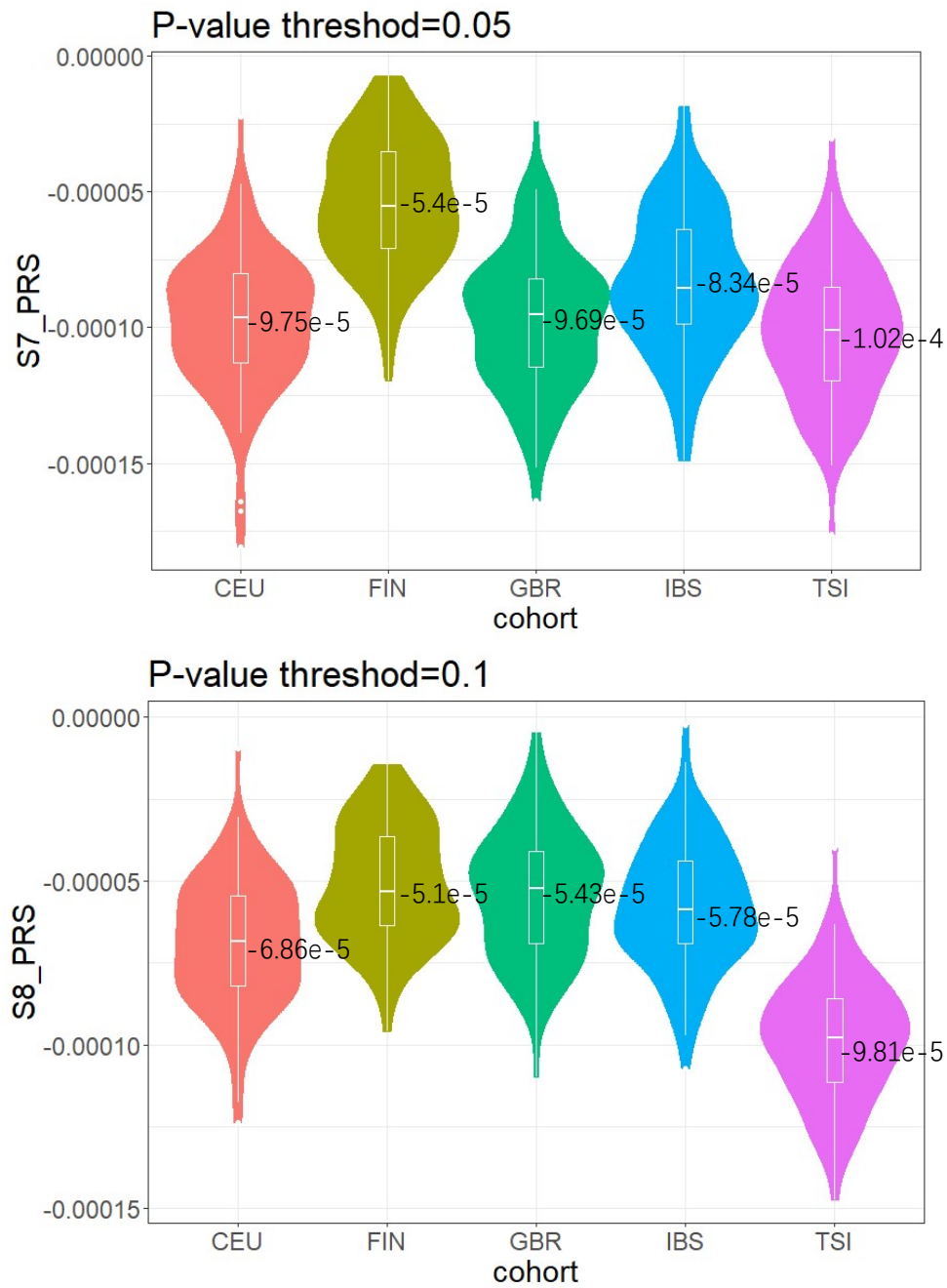


Fig. 2 Polygenic risk scores for 5 European cohorts at 8 PTs

All 5 cohorts genotype are from 1000 Genomes Project Phase 3 populations of a total of 2,458,861 variants. PRSs are aggregate weighted effect sizes, and effect sizes were measured by \pm diopeters per effect allele, myopia is determined by spherical equivalent < -0.50 diopeters, so minus value of PRS stands for towards myopia, and plus means against myopia.

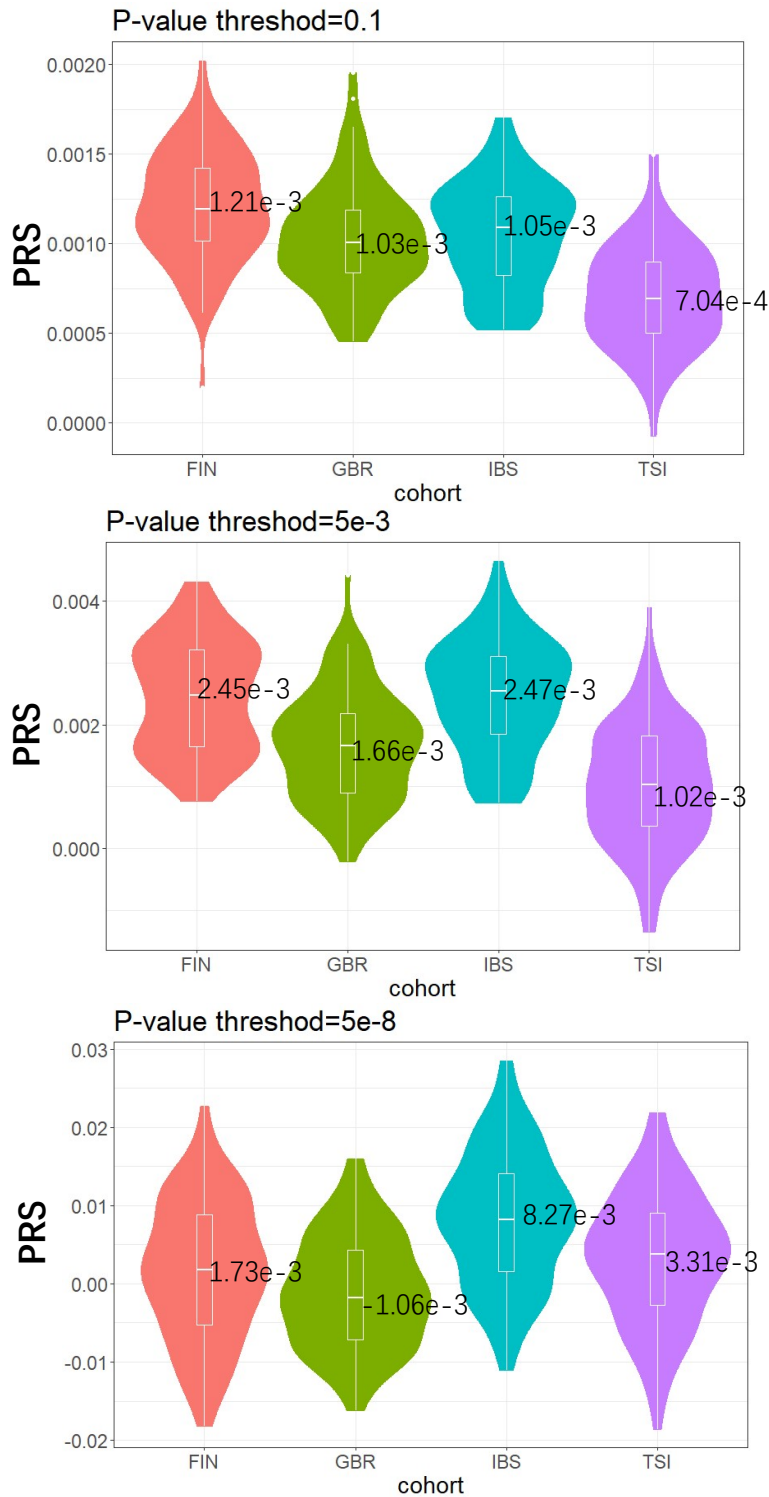


Fig. 3 Polygenic risk scores replication for 4 European cohorts at 3 PTs

All 4 cohorts genotype are from 1000 Genomes Project Phase 3 populations of a total of 80,855,702 variants. PRSs are summation of weighted effect sizes, and effect sizes were measured by \pm diopeters per effect allele, myopia is determined by spherical equivalent <-0.50 diopeters, so minus value of PRS stands for towards myopia, and plus means against myopia.

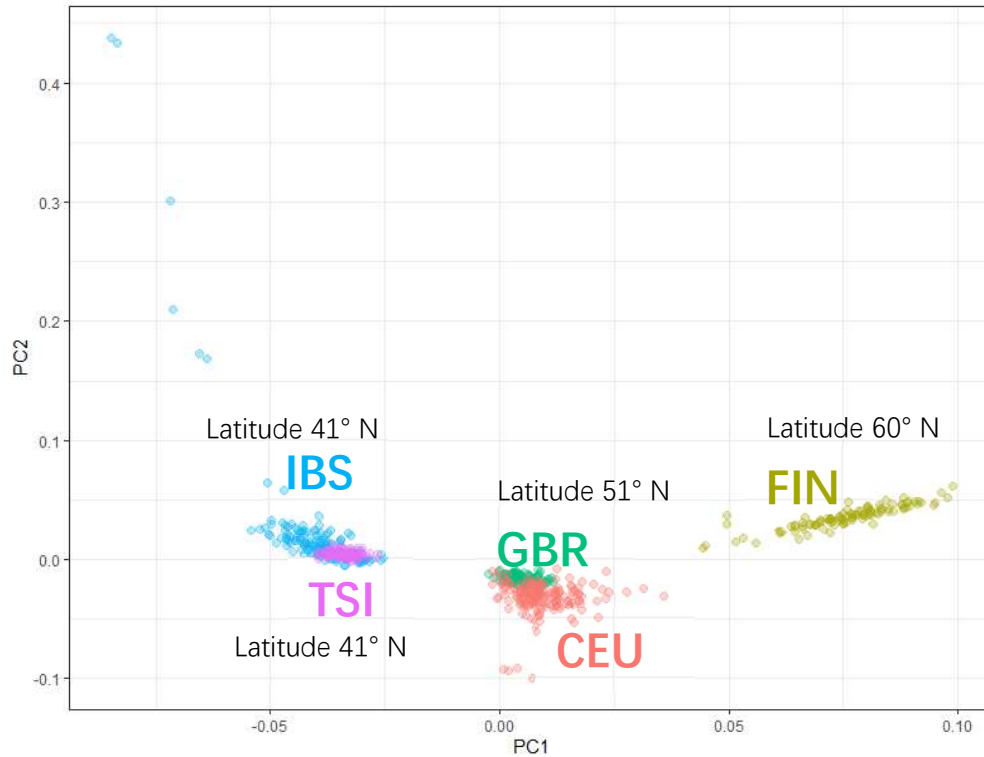


Fig. 4 Principal component analyses for ~90,000 scored variants of the 5 European cohorts

Scored variants (Table 1) of 5 cohorts in target data 1 were analyzed by GCTA (Yang et al. 2011), ratio of PC1 and PC2 are 38.69% and 24.23%. Notice latitudes distribution of each population's capital city are similar to PC1.

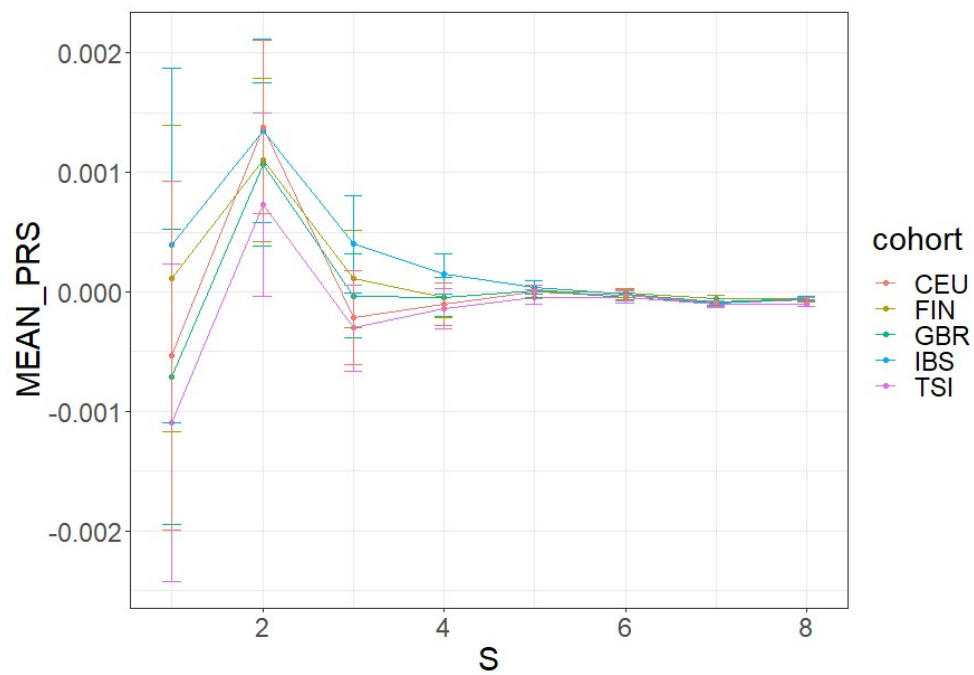


Fig. 5 PRS distribution at 8 PTs for target data 1

PRSs mean and standard error distribution of S1–S8 PTs (Table 2) for all 5 cohorts. S8 has the most variants and S1 has the least, the PRSs drastically deviated from each other when variants number gets smaller.

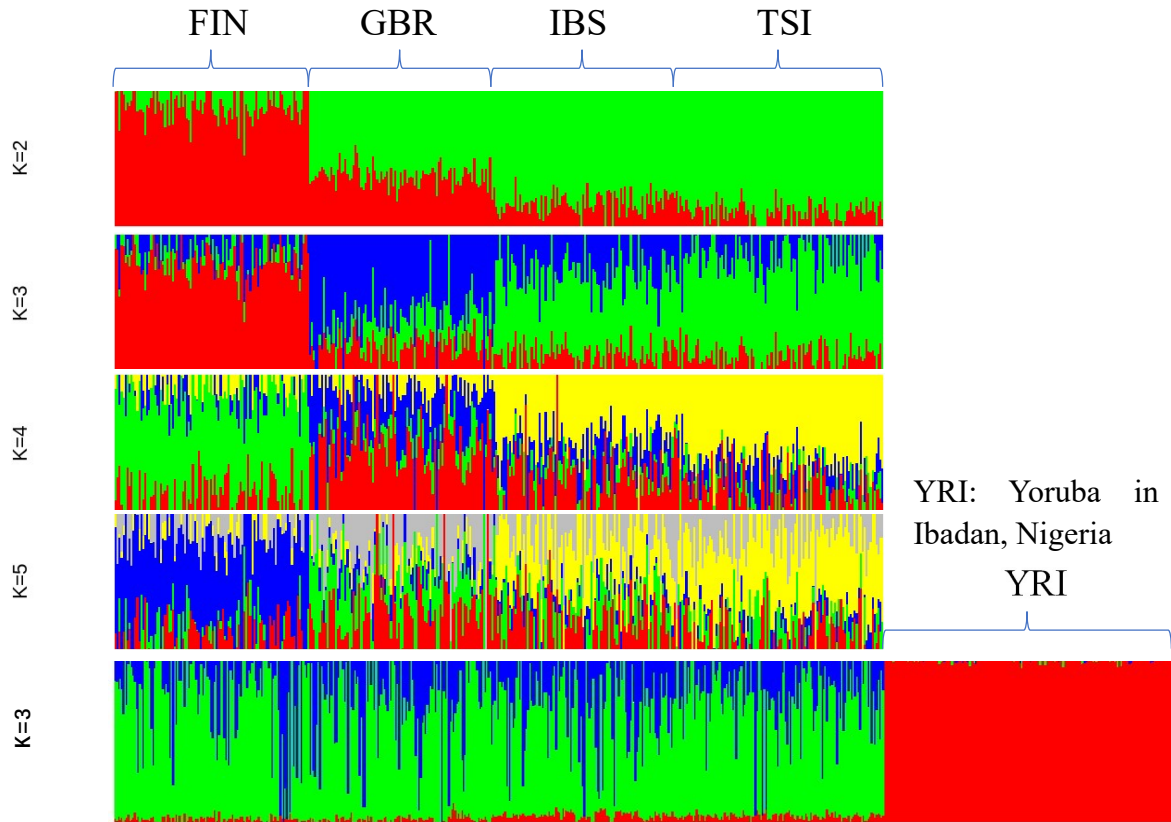


Fig. 6 Ancestry proportion analysis of target data 1

Ancestry proportions by ADMIXTURE analysis (Alexander DH et al. 2011) inputting base data and target data 1, at $PT=5 \times 10^{-3}$ extracting 10,253 scored variants. YRI is compared as outer group.

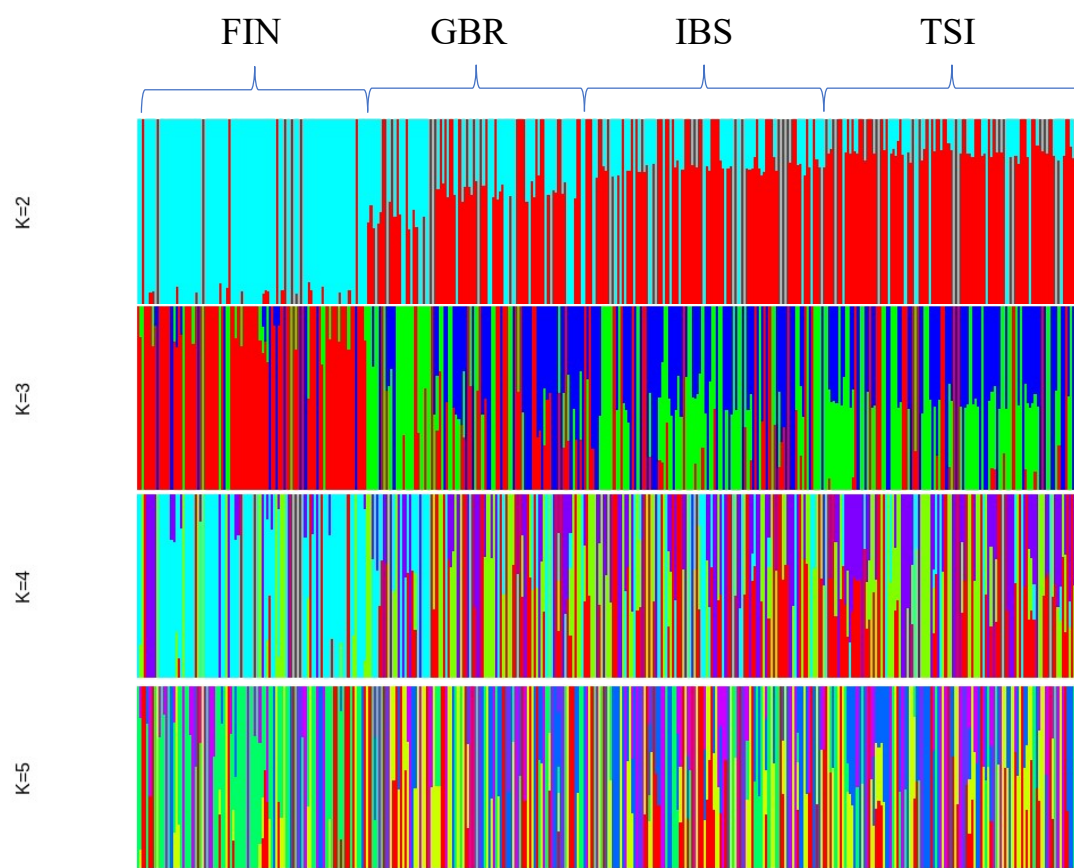
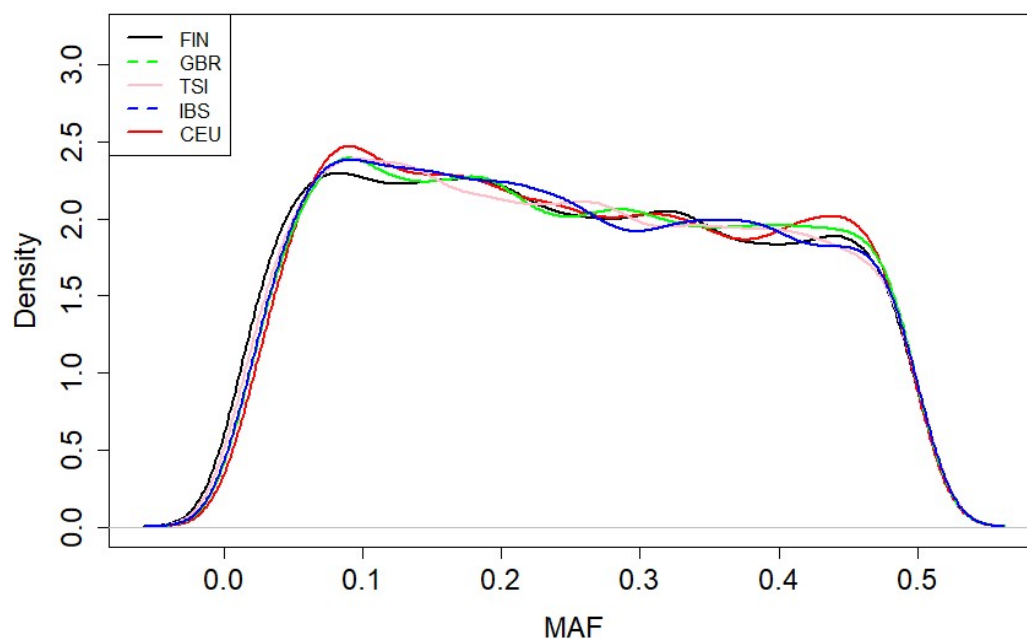


Fig. 7 Ancestry proportion analysis of target data 2

Ancestry proportions by ADMIXTURE analysis (Alexander DH et al. 2011) inputting base data and target data 2 for all 569,974 scored variants.

Target data 1 MAF density, PT = 5e-3



Target data 2 MAF density, PT = 5e-3

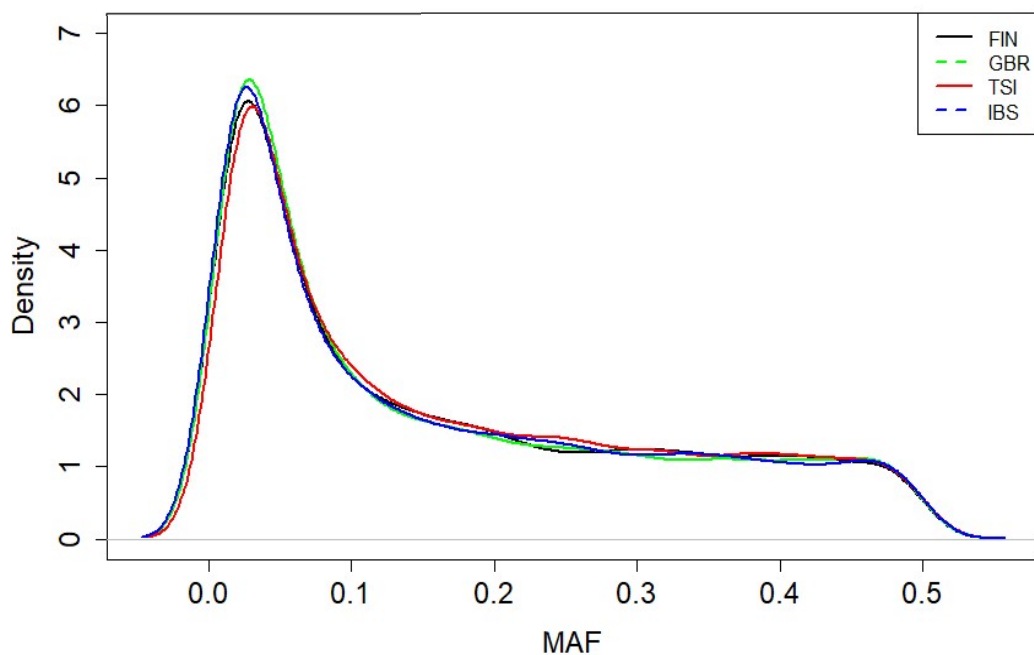
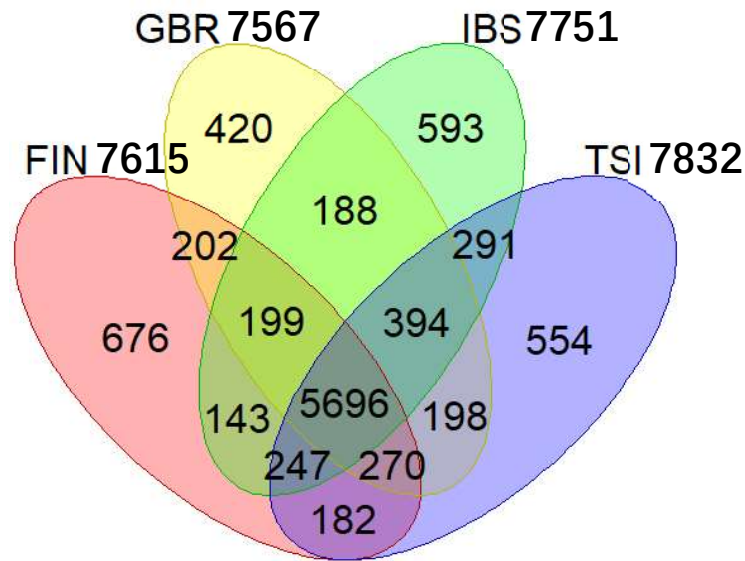


Fig. 8 Variant minor allele frequency (MAF) distribution of target data 1 and target data 2

MAF density distribution at $PT = 5 \times 10^{-3}$ of target data 1 and target data 2, notice in target data 2 there are more rare alleles (MAF < 0.1).

A.



B.

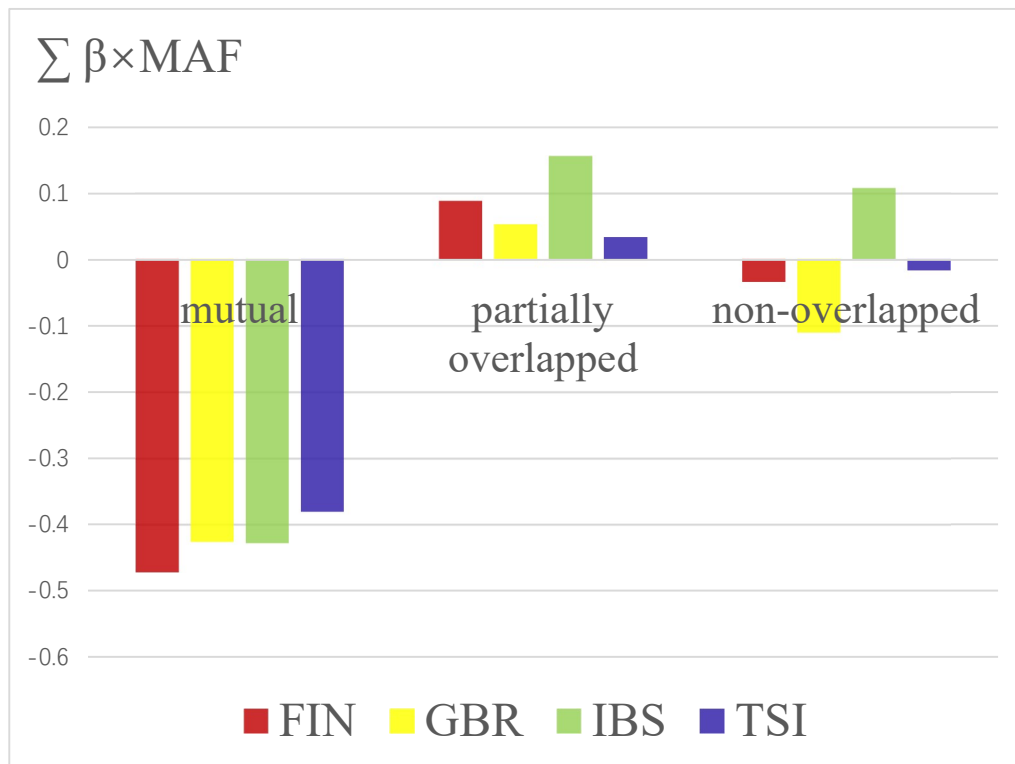
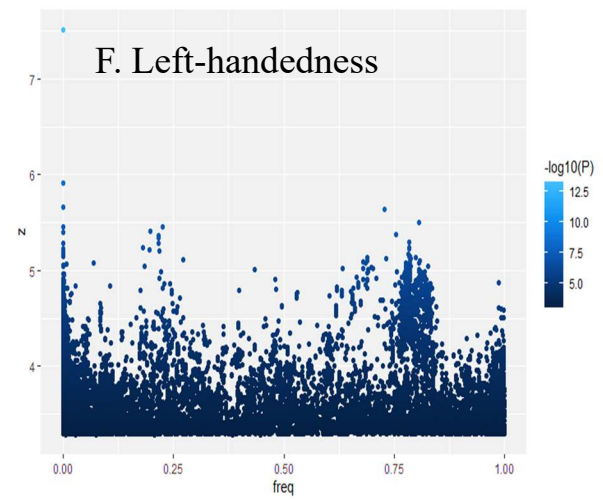
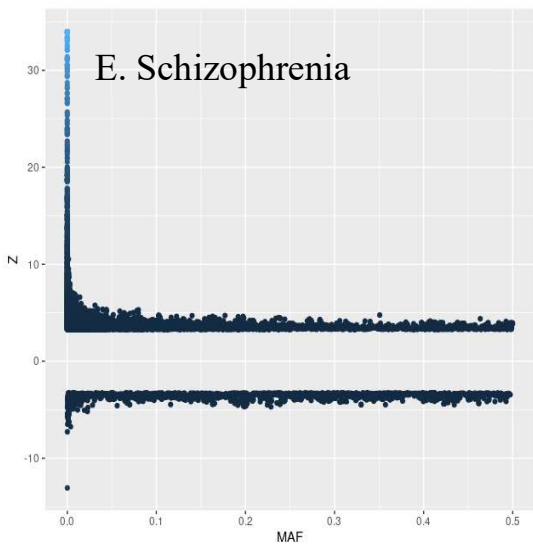
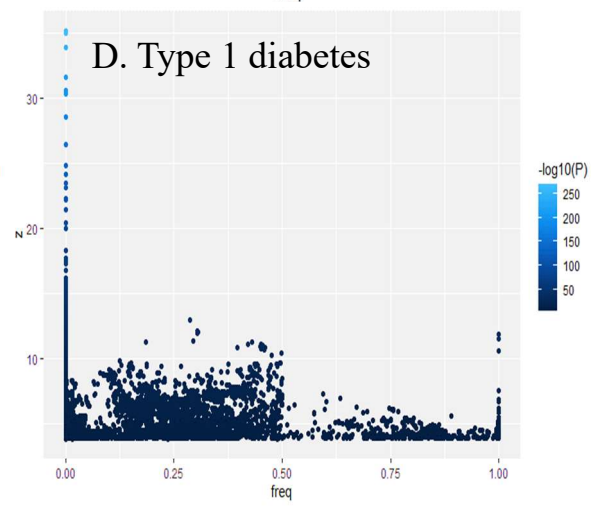
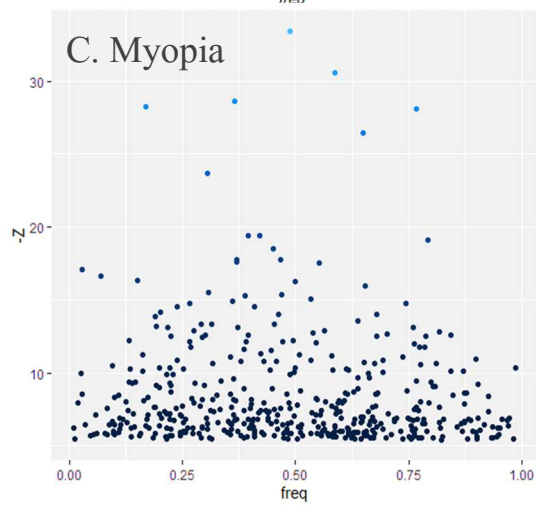
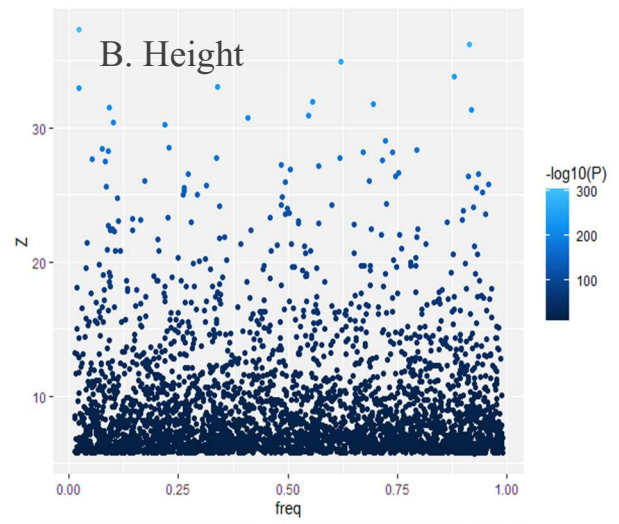
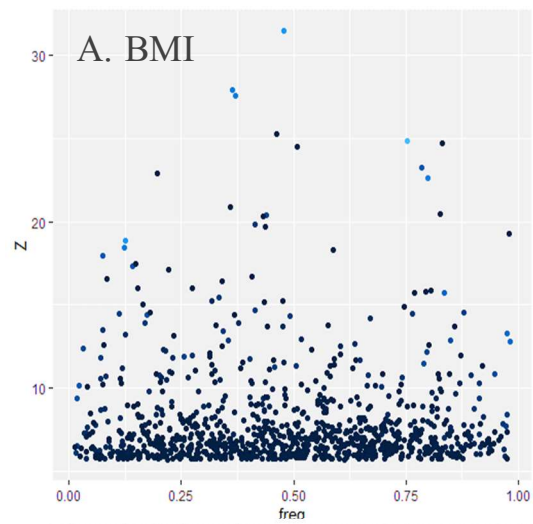


Fig. 9 Myopia associated variants comparison

A: Scored variants at P value $< 5 \times 10^{-3}$ of target data 1, Venn diagram by R.

B: Variants categorized in Venn diagram analysis by 5,696 mutual variants, 2,244 non-overlapped variants and rest as partially overlapped variants. Accumulated $\beta \times \text{MAF}$ of each variant in each cohort revealing peculiar genetic impact in non-overlapped variants in IBS.



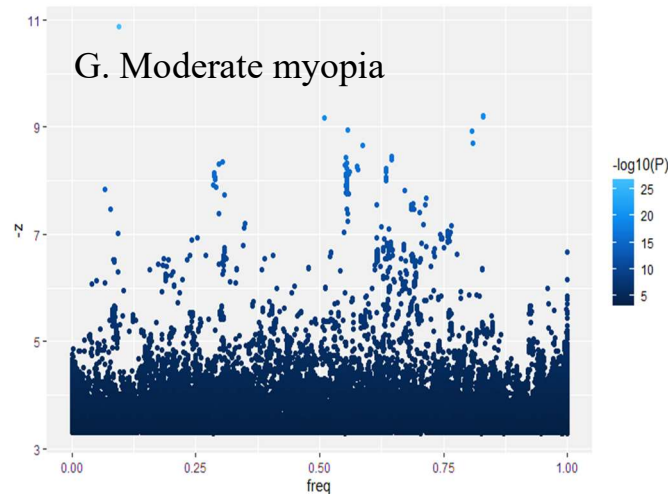


Fig. 10 Effect size-frequency distribution across different traits

A, B: BMI and height risk allele distribution improvised from summary statistics of a Meta-analysis of GWAS for height and body mass index in ~700,000 individuals of European ancestry (Yengo L, et.al. 2018).

C: Myopia risk allele distribution improvised from GWAS summary statistics (P.G. Hysi et al. 2020).

D, E, F, G: Type-1 Diabetes, schizophrenia, left-handedness and moderate myopia risk allele distribution improvised from UK GWAS round 2 summary statistics (<http://www.nealelab.is/uk-biobank/>).

Effect sizes are calculated by Z , which equals to β/SE . P value indicating significant of each allele, and more significant allele often has stronger effect size. Frequencies of traits under stronger selective pressure (D, E) tend to be rare, and traits under weak selective pressure tend to have more higher MAF (A, C, G), and evenly distributed frequency traits are under moderate selective pressure (B, F).