Statistical methods in genome-wide association studies

Erna Valdís Ívarsdóttir

Faculty of Physical Sciences
University of Iceland
2020

Statistical methods in genome-wide association studies

Erna Valdís Ívarsdóttir

Dissertation submitted in partial fulfillment of a
Philosophiae Doctor degree in Statistics

PhD Committee
Dr. Daníel F. Guðbjartsson
Dr. Gunnar Stefánsson
Dr. Bjarni V. Halldórsson

Opponents
Dr. Mark Daly
Dr. Thor Aspelund

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, November 2020

Statistical methods in genome-wide association studies
Dissertation submitted in partial fulfillment of a *Philosophiae Doctor* degree in Statistics

Faculty of Physical Sciences
School of Engineering and Natural Sciences
University of Iceland
Dunhagi 5
107, Reykjavik
Iceland

Telephone: 525 4000

ISBN 978-9935-9564-2-2

Printed in Iceland by Háskólaprent
Reykjavik, Iceland, November 2020

Abstract

The aim of genome-wide association studies (GWAS) is to identify sequence variants that influence human traits or diseases. Previous GWAS have mostly focused on finding variants that affect the mean of a trait or disease risk under an additive model. However, variants can contribute to traits in different ways, such as under a recessive mode of inheritance and by affecting the variance of quantitative traits. In this thesis we use different statistical models to detect variants associating with sensory traits and explore relationships between correlated phenotypes. Furthermore, we implement a variance model to detect sequence variants that affect the variance of quantitative traits and we explore the effect of variants on the variance of glucose levels.

In Paper I we estimate the effect of 36 glucose variants on the between subject and within subject variance of glucose levels. We found that some variants that affect the mean also affect the variance. The trend was that variants that increased mean and between subject variance of fasting glucose increased type 2 diabetes (T2D) risk, while variants that increase the mean but reduce the variance do not. We found that the effect of variants on the between subject variance of glucose levels are as important for genetic risk prediction of T2D as the effect of variants on the mean. Furthermore, the variants that increased between subject variance created correlation between close relatives and will thus increase heritability estimates.

In Paper II we conduct a GWAS on structural measures of the corneal endothelium that are used in clinic to evaluate the health of the cornea. We detected associations at 7 novel loci, one of which is an intergenic variant near *ANAPC1* that strongly associates with decreased endothelial cell density and accounts for a quarter of the population variance of cell density. The variant near *ANAPC1* does not affect risk of corneal diseases or glaucoma in our data, which shows that even though low endothelial cell density is associated with ocular diseases, low cell density does not in and of itself lead to the development of disease.

In Paper III we conduct GWAS meta-analysis of age-related hearing impairment (ARHI) using both the additive and recessive models. Previous GWAS on ARHI have reported common variants with small to moderate effects, while in this study, 13 of the 21 novel variants have rare genotypes with large effects. Six of the novel variants associate with ARHI under the recessive model, some of which would not have been detected under the additive model. We constructed an ARHI genetic risk score (GRS) using common variants and show that individuals in the top GRS decile develop ARHI 10 years earlier than those in the bottom decile, and their risk of ARHI is comparable to carriers of rare highly penetrant ARHI variants while the rare ARHI variants predispose to more severe ARHI than the common variants.

Our findings shed a new light on the genetics of glycemic traits, the corneal endothelium and ARHI and highlight the importance of applying different statistical models when analyzing the effects of variants on phenotypes.

Útdráttur

Markmið víðtækra erfðamengisleita er að finna erfðabreytileika sem hafa áhrif á mannlega eiginleika og sjúkdóma. Hingað til hafa víðtækar erfðamengisleitir lagt áherslu á að finna erfðabreytileika sem hafa áhrif á meðaltal mælanlegra eiginleika eða áhættu á sjúkdómum með því að nota samleggjandi líkan. Breytileiki í erfðamenginu getur einnig haft áhrif á mannlega eiginleika á mismunandi hátt, til dæmis með víkjandi hætti eða með því að hafa áhrif á dreifni mælanlegra eiginleika í stað meðaltals. Í þessari ritgerð er notast við mismunandi tölfræðilíkön til að finna erfðabreytileika sem hafa áhrif á eiginleika hornhimnunnar og heyrn, ásamt því að rannsaka orsakasambönd milli tengdra svipgerða. Einnig eru útfærð dreifnilíkön til að finna sambönd milli erfðabreytileika og dreifni mælanlegra gilda og skoðuð áhrif erfðabreytileika á dreifni glúkósa í blóði.

Í grein I voru metin áhrif 36 þekktra glúkósa erfðabreytileika á dreifni glúkósamælinga milli einstaklinga og innan einstaklinga. Í ljós kom að sumir erfðabreytileikar sem hafa áhrif á meðaltal glúkósa hafa einnig áhrif á dreifnina. Einnig sást að erfðabreytileikar sem hafa áhrif á aukið meðaltal af glúkósa og aukna dreifni milli einstaklinga auka einnig líkur á sykursýki, á meðan þeir sem juku meðtalið en drógu úr dreifni hafa ekki áhrif á sykursýki. Einnig var sýnt fram á að erfðabreytileikar sem auka dreifni glúkósa milli einstaklinga búa til fylgni milli skyldra einstaklinga og hafa þar af leiðandi áhrif á mat á arfgengni.

Í grein II var framkvæmd víðtæk erfðamengisleit fyrir hornhimnumælingar og fundust 7 erfðabreytileika sem voru áður óþekktir. Einn af þeim er erfðabreytleiki nálægt *ANAPC1* sem hefur veruleg áhrif á frumuþéttleika í innþekju hornhimnunnar og útskýrir fjórðung af heildardreifni frumuþéttleika í þýðinu. Þessi erfðabreytileiki hefur hinsvegar ekki áhrif á hornhimnusjúkdóma eða gláku, sem sýnir að þrátt fyrir fylgni milli frumuþéttleika innþekjunnar og augnsjúkdóma, þá veldur lítill frumuþéttleiki ekki auknum líkum á þessum augnsjúkdómum.

Í grein III var framkvæmd safnrannsókn víðtækra erfðamengisleita á aldurstengdri heyrnarskerðingu þar sem notað voru bæði samleggjandi og víkjandi líkön. 21 áður óþekktir erfðabreytileikar fundust, þar af 13 sjaldgæfir. Sex af áður óþekktu erfðabreytileikunum hafa áhrif á aldurstengda heyrnarskerðingu með víkjandi hætti. Reiknað var fjölgena áhættuskor út frá algengu erfðabreytileikunum og sýnt að einstaklingar í efstu tíund áhættuskorsins fá aldurstengda heyrnarskerðingu að meðaltali 10 árum fyrr en þeir sem eru í neðstu tíund áhættuskorsins. Einnig sást að áhætta þeirra sem eru í efstu tíund áhættuskorsins er sambærileg áhættu þeirra sem hafa sjaldgæfar stökkbreytingar sem valda aldurstengdri heyrnarskerðingu en sjaldgæfu erfðabreytileikarnir valda þó verri heyrnarskerðingu en þeir algengu.

Niðurstöður þessara rannsókna varpa nýju ljósi á erfðafræði glúkósa í blóði, innþekju hornhimnunnar og aldurstengdrar heyrnarskerðingar, ásamt því að sýna fram á mikilvægi þess að nota fjölbreytt tölfræðilíkön til að meta áhrif erfðabreytileika á svipgerðir.

Table of Contents

Abstract .... **iii**
List of Figures .... **ix**
List of Tables .... **xii**
Abbreviations .... **xiii**
List of publications .... **xiv**
Acknowledgements .... **xv**
Part I - Thesis **1**
1 Introduction **1**
2 Background **3**
2.1 Genetics 3
2.2 GWAS models 4
2.2.1 Inheritance models 5
2.2.2 Genetic effect on the variance of traits 6
2.3 Genetic epidemiology 7
2.3.1 Heritability 7
2.3.2 Mendelian randomization methods 8
2.4 Phenotypes 9
2.4.1 Glucose levels and type 2 diabetes 9
2.4.2 The cornea 9
2.4.3 Hearing 11
3 Aim **13**
4 Materials and Methods **15**
4.1 Study Subjects 15
4.1.1 The deCODE health study 15
4.1.2 Paper I 15
4.1.3 Paper II 15
4.1.4 Paper III 15
4.1.5 Ethical statement 16
4.2 Genotypic data 16
4.2.1 Iceland 16
4.2.2 UK Biobank 16
4.3 Association testing 17
4.3.1 Binary traits 17
4.3.2 Quantitative traits 17
4.3.3 Variance model 18
4.3.4 Meta-analysis 21
4.3.5 Adjusting for population stratification 21
4.4 Genetic effect on variance and heritability 21
4.5 Correlation between effect sizes 23

4.6 Genetic risk score 24

5 Summary of key results 25

5.1 Paper I - Sequence variants affecting the variance of glucose levels 25

5.1.1 Performance of the variance models 25

5.1.2 Data summary 27

5.1.3 Effect of glucose variants on variance in glucose levels 29

5.1.4 Replication of variance effects 30

5.1.5 Effect on glucose levels vs. effect on T2D risk 30

5.1.6 Possible interaction 31

5.1.7 Genetic risk score 31

5.1.8 Heritability 32

5.2 Paper II - A GWAS on structural measures of the corneal endothelium 32

5.2.1 Sex and age effect on corneal traits 32

5.2.2 Study design 33

5.2.3 GWAS results 34

5.2.4 The effect of cell density associating variants on diseases and other corneal metrics 36

5.2.5 Glaucoma and corneal measures 36

5.3 Paper III – A GWAS meta-analysis of age-related hearing impairment 38

5.3.1 Demographics of ARHI in Iceland 38

5.3.2 Study Design 39

5.3.3 GWAS results 41

5.3.4 Genetic risk scores predict ARHI risk 44

5.3.5 Association of ARHI variants with ARHI under the full genotypic model 46

5.3.6 ARHI and tinnitus 46

6 Discussion 49

6.1 Conclusions 49

6.1.1 Paper I - Sequence variants affecting the variance of glucose levels 49

6.1.2 Paper II - A GWAS on structural measures of the corneal endothelium 49

6.1.3 Paper III – A GWAS meta-analysis of ARHI 50

6.2 Future perspectives 50

6.2.1 Paper I - Sequence variants affecting the variance of glucose levels 50

6.2.2 Papers II and III – GWAS on corneal traits and age-related hearing impairment 51

References 53

Part II - Papers 59

Paper I 61

Paper II 67

Paper III 79

List of Figures

Figure 1. *Coding of genotypes under different models. The blue color represents alleles on the paternally inherited chromosome, and orange on the maternally inherited chromosome.* A *represents the minor allele and* a *the more common allele.* 5

Figure 2. *Examples of sequence variants affecting the variance of a quantitative trait. The grey solid line shows the density of a trait for non-carriers and the grey dotted line shows the mean. The orange and red lines show the same for heterozygotes and homozygotes carriers, respectively. The examples are for a) a variant that associates with both increased mean and variance of a quantitative trait and b) a variant that associates with increased variance but does not affect the mean.* 6

Figure 3. *A schematic figure of the cornea.* 10

Figure 4. *Images of the corneal endothelium obtained with a specular microscopy.* 10

Figure 5. *The power of detecting BS variance effect for different simulation scenarios using the BS variance model (round dots) and the Levene's test (triangles) for different true variance effects (between 1.01 and 1.06) and different sample sizes. Figure a) shows the results for the simulated phenotype using sample sizes from 10,000 to 100,000 individuals and b) shows additionally the sample sizes from 200,000 to 500,000 individuals.* 25

Figure 6. *The fraction of significant BS variance effects for a simulated normal phenotype for 100,000 individuals using different mean effects (between 0.1 and 0.7 SD) and different minor allele frequency (MAF) of the simulated variants. The fraction was computed as the number of BS variance associations with P<0.05 divided by the total number of tests across 1000 simulations for each simulation scenario.* 26

Figure 7. *a) The distribution of the simulated number of measurements per person. b) The power of detecting WS variance effect for different simulation scenarios using the WS variance model (round dots) and the Levene's test (triangles) for different true WS variance effects (between 1.01 and 1.06) and different sample sizes.* 27

Figure 8. *A histogram that shows the number of all glucose measurements per person.* 28

Figure 9. *Mean and variance of glucose levels against age. (a) The average and (b) the variance of fasting and not fasting glucose (mmol/L) levels per age group for both sexes.* 28

Figure 10. *Fasting glucose (FG) GRS based on the 36 fasting glucose variants.* 31

Figure 11. *The mean of the trait by age groups for women in red and men in blue. The gray lines show 95% confidence intervals.* 33

Figure 12. Manhattan plots showing the association results for the GWAS of the four corneal traits. The $-log_{10}$ P-values are shown for each variant against their chromosomal position. 34

Figure 13. The mean cell density per age group for non-carriers of the ANAPC1 variant in grey, heterozygous carriers in yellow and homozygous carriers in green. 35

Figure 14. Effect of previously reported glaucoma variants on corneal traits for the glaucoma risk increasing allele. Effects on the traits are shown for significant associations after adjusting for multiple testing with the Benjamini–Hochberg false discovery rate procedure for each variant. Effect are shown in red for positive effects and blue for negative effects. 37

Figure 15. Histograms of PTA hearing thresholds from a) DHS and b) NIHSI datasets. 38

Figure 16. The prevalence of hearing impairment by age groups in the DHS dataset. Shown are the fraction of women (squares) and men (round dots) per age group with mild in grey, moderate in yellow and severe hearing impairment in red. 39

Figure 17. A flowchart explaining the study design of the ARHI GWAS meta-analysis. The UK Biobank (UKB) GWAS was performed on two genotype datasets labelled in red and blue. WGS = Whole genome sequenced, HRC = Haplotype Reference Consortium, WES = Whole exome sequenced, PTA = Pure tone average, DHS = deCODE health study, NIHSI = National Institute of Hearing and Speech in Iceland. 40

Figure 18. Manhattan plots showing the association results for the ARHI meta-analysis under the a) the additive model and b) the recessive model. The $-log_{10}$ P-values are shown for each variant against their chromosomal position. 40

Figure 19. A flowchart showing a summary of the results for the ARHI GWAS meta-analysis. 42

Figure 20. The cumulative risk of mild, moderate and severe hearing impairment in the DHS dataset among the 4.9% of subjects that are carriers of any of the 16 rare ARHI variants (round dots) and the 95% that are not carriers (squares). 43

Figure 21. Changes in PTA hearing thresholds by age for carriers of the variants in a) LOXHD1 and b) FBF1. In figure a, PTA means of non-carriers in the DHS dataset are represented with grey dots, means of heterozygotes in DHS dataset with orange dots and the PTA hearing threshold of the homozygous carriers in DHS and NIHSI datasets are indicated by red squares. In figures b PTA means of non-carriers in the DHS dataset are represented with grey dots and the PTA hearing threshold of the *heterozygous carriers in DHS and NIHSI datasets are indicated by yellow squares.* 44

Figure 22. a) The ORs for ARHI for each GRS decile in the DHS and NIHSI datasets relative to the bottom decile. b) The cumulative risk of ARHI among subjects in the DHS dataset in the bottom GRS decile in blue and the top GRS decile in red. 45

Figure 23 The effect of the ARHI variants on ARHI is plotted against their effect on tinnitus for a) all ARHI variants and b) zoomed-in on variants with ARHI OR <1.35. ARHI variants detected under the additive model are colored blue and ARHI variants detected under the recessive model are colored red. Variants that affect tinnitus, controlling the false discovery rate at 0.05, are plotted with darker color and labelled with their corresponding gene. The effects are shown for the ARHI risk increasing allele. Error bars represent 95% confidence intervals. The dotted lines represent results from a weighted linear regression using MAF(1-MAF) as weights, red for recessive variants and blue for additive, and the weighted correlation coefficients (r) and the corresponding P-values are shown in figure a. 47

List of Tables

Table 1. *The FPR of the BS variance test and Levene's test for different phenotype distributions estimated from 1000 simulations for 50,000 individuals.* 26

Table 2. *The association of three glucose variants on the BS variance of glucose levels.* 29

Table 3. *The association of three glucose variants on the WS variance of glucose levels.* 29

Table 4. *The association of glucose variants on the BS and WS variance of glucose levels in the Icelandic and Iranian datasets.* 30

Table 5. *The association of the TCF4 and ANAPC1 variants with ocular diseases in a meta-analysis of GWAS results from Iceland and the UK Biobank. The number of cases and controls are shown for each disease.* 36

Table 6. *The first two columns show the effect of glaucoma status on corneal traits, adjusting for age and sex. The other two columns show the correlation (R^2) between the effect of glaucoma variants on glaucoma and their effect on the corneal traits.* 37

Table 7. *The effect of seven ARHI variants on ARHI per genotype, four that had stronger effect on homozygous carriers than expected by the additive model, and three recessive variants that have been reported under the additive model. The deviation column shows the P-value when comparing the full genotypic model to the additive model.* 46

Abbreviations

ARHI	Age-related hearing impairment
BS	Between subjects
CCT	Central corneal thickness
CV	Coefficient of cell size variation
dB HL	Decibel hearing level
DHS	deCODE health study
DNA	Deoxyribonucleic acid
FECD	Fuchs corneal endothelial dystrophy
GRS	Genetic risk score
GWAS	Genome-wide association study
HEX	Percentage of hexagonally shaped cells
Hz	Hertz
LD	Linkage disequilibrium
LOF	Loss-of-function
MAF	Minor allele frequency
MCD	Macular corneal dystrophy
MR	Mendelian Randomization
NIHSI	National hearing and speech institute of Iceland
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
UKB	UK Biobank
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WS	Within subjects

List of publications

This thesis is based on the following three papers, which will be referred to in the text by their Roman numbers.

I. Effect of sequence variants on variance in glucose levels predicts type 2 diabetes risk and accounts for heritability.

II. Sequence variation at *ANAPC1* accounts for 24% of the variability in corneal endothelial cell density.

III. The genetic architecture of age-related hearing impairment.

Acknowledgements

This doctoral study was conducted at deCODE genetics and I would like to thank deCODE for giving me the opportunity to work on exciting research and to further my studies. I am sincerely grateful for being able to learn from the incredibly talented scientists that work at deCODE, as well for the kindness and support I have received from them.

Especially, I would like to express my gratitude to my principal supervisor, Daníel Guðbjartsson, for his guidance and support through the years. I would also like to thank all the co-authors of the appended papers for their essential contributions and my doctoral committee, Gunnar Stefánsson and Bjarni Halldórsson. Special thanks to Kári Stefánsson for his supervision and for creating the unique research environment that deCODE is.

I am extremely fortunate to be surrounded by people who always support and encourage me. I am especially thankful for my parents for their endless love and support. I also want to thank the rest of my family and my dearest friends, for everything they have done for me. I also feel very lucky to have made great friends at deCODE and I especially want to thank Stefanía and Guðný for their proofreading and for always being there for me. I am also grateful to my late husband who encouraged me to choose this field of study.

Finally, I want to thank Bjössi for his kindness, love and support and my wonderful daughter Ásta for making me smile every single day.

Part I - Thesis

1 Introduction

Genome-wide association studies (GWAS) aim to identify association between phenotypes and sequence variation in the genome. Understanding the phenotypic effects of sequence variation can give insight into how to prevent and treatment of diseases.

GWAS based on chips containing hundreds of thousands of single nucleotide polymorphisms have transformed the study of human genetics and thousands of genotype-phenotype associations have been discovered. Advances in sequencing technologies have considerably reduced the cost of whole genome sequencing human genomes and allowed sequencing to be performed on a large scale. This has resulted in a rapid increase in the number of individuals with their whole genomes sequenced, allowing the identification of most of their sequence variation and discovery of associations with rare variants with large effects. Currently, around 50,000 Icelanders have been whole genome sequenced by deCODE genetics which allowed the identification of over 34 million high quality sequence variants that have been imputed into over 110,000 additional chip typed Icelanders[1,2].

Vast amount of phenotypic data is available in the deCODE database. A recent addition of phenotypes is being obtained in the deCODE health study, an ongoing population-based study, which involves a comprehensive phenotyping of the recruited subjects. Measurements include audiometric tests and various ocular measures which have not been analyzed before in large GWAS.

Sequence variants can affect phenotypes in different ways and we therefore need to consider what statistical model is best suited to find genotype-phenotype associations. The additive model is easy to interpret and is therefore the most commonly used model, but other models include the recessive, dominant, parent-of-origin and full genotypic models. Most quantitative trait GWAS to date have focused on discovering sequence variants that affect the trait mean. However, variants can also affect the variability of quantitative traits, both within-subject variance and between-subject variance. Assessing these variance effects may lead to the discovery of novel genotype-phenotype associations and has the potential of improving our understanding of previously associated sequence variants.

This thesis is based on the three papers listed in the Abstract. The goals of this research can be divided into two main categories:

1. Implementing a variance model to test for an association between sequence variants and the variance of glucose levels and investigate how the effect of these variants on variance affect heritability estimates (Paper I).
2. Using different statistical models to search for sequence variants associating with sensory measurements (auditory and ocular) obtained from the deCODE health study and exploring the causal relationship between correlated traits and diseases (Papers II and III).

This thesis consists of two parts. Part I contains introduction and a summary of the three papers and Part II contains two published peer-reviewed journal papers (Paper I and Paper II) and a submitted paper (Paper III). In part I, chapter 2 provides an overview of the genetic, phenotypic and statistical background for this thesis. The specific aims of this project are listed in Chapter 3 and Chapter 4 contains a summary of the materials and methods used in these studies. In Chapter 5 the main results from the three papers are summarized and in Chapter 6 the results are discussed.

2 Background

2.1 Genetics

Human cells have 23 pairs of chromosomes that carry the individual's genetic information. The chromosomes are made of histones and deoxyribonucleic acid (DNA) molecules. The genetic information is carried by the DNA which is a long ladder-like macromolecule that twists to form a double helix made up of four nucleobases; adenine (A), thymine (T), cytosine (C) and guanine (G). The nucleobases bind together to make base pairs, A with T and C with G. The genetic information is then coded in the linear sequence of the bases. DNA is decoded to make ribonucleic acid (RNA) in a process call transcription. Genes are segments of DNA that serve as a template for making a functionally important RNA molecule which is used to make polypeptides that form proteins in a process called translation. Examples of proteins are enzymes, antibodies, structural components and hormones which have various different roles in the body.

A sequence variant is defined as a DNA locus with at least one base being different between individuals. The simplest form of a sequence variant is when a single base varies and is called a single-nucleotide polymorphisms (SNP). Different variations of sequence variants are known as alleles. SNPs have two alleles, and the allele that is less common is called a minor allele and the other is called a major allele. The minor allele frequency (MAF) of a sequence variant is then the frequency of the minor allele in the population. A genotype of an individual is the alleles of the sequence variant that the individual carries. For instance, if we use A to represent the minor allele and a to represent the major allele in the population, there are four possible genotypes, aa, Aa, aA and AA, where the first letter represents the paternally inherited allele and the second letter represents the maternally inherited allele. An individual with genotype aa is a homozygous non-carrier of the minor allele. Individuals with genotype Aa or aA are heterozygous carrier of the minor allele, while an individual with genotype AA is a homozygous carrier of the minor allele. The genotype is usually coded as the allele count; 0 for non-carriers, 1 for heterozygotes and 2 for homozygotes.

Sequence variants that are located close together can be correlated because during meiosis, alleles on the same chromosome are inherited together, except when a crossover occurs. The crossover is often called recombination. The non-random association of alleles at different loci is called linkage-disequilibrium (LD). If p_{AB} is the frequency of individuals carrying the pair of alleles A and B, p_A is the frequency of A and p_B is the frequency of B, then two different measures of LD are

$$D' = (p_{AB} - p_A p_B) \,/\, D_{max},$$

where $D_{max} = min\{p_A(1 - p_B), p_B(1 - p_A)\}$ if $D > 0$ and $D_{max} = min\{p_A p_B, (1 - p_A)(1 - p_B)$ if $D < 0$, and

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A(1-p_A)p_B(1-p_B)}.$$

The correlation, r^2, is the more commonly used measure of LD because it is a convenient measure to identify variants that have information about the effects of other variants due to correlation, often called tagging variants.

When the genotypes of a sequence variant are not known, we use imputed genotype probabilities which are calculated using knowledge about LD between variants and haplotype sharing between individuals[1,2].

Genetic variation in humans ranges from SNPs to loss or gain of whole chromosomes. Other types of variants are for example deletions, insertions, duplications, inversions and short tandemly repeated sequences called microsatellites. Genes have regions that code for amino acids called exons and regions between exons that are removed after transcription called introns. Protein-coding regions account for less than 2% of the whole genome. Sequence variants located in protein-coding regions are called coding variants and they are categorized by the effect they have on the translated amino acids. For example, when a SNP in a coding region alters the amino acid, it is called a missense variant. A variant that results in reduced or complete loss of protein function is called a loss-of-function (LOF) variant. Examples of LOF variants are stop gained variants, which result in a premature termination codon, and frameshift variants, which shift the way a sequence is read. Coding variants are more likely to affect phenotypes than intronic or intergenic variants.

2.2 GWAS models

GWAS aims to identify association between sequence variation in the genome and an observable trait or disease of interest. Through the whole-genome sequencing (WGS) of around 50,000 Icelanders, over 34 million high quality sequence variants have been identified in the population. When conducting a GWAS, for each variant a statistical test is performed to see if the variant is affecting the trait of interest. When analyzing millions of variants, correcting for multiple testing is necessary and a Bonferroni corrected P-value threshold for one million independent tests (5×10^{-8}) is commonly used in GWAS assigning equal prior probability to all variants. However, because variants that are in coding regions are more likely to affect phenotypes, using a weighted Bonferroni method where genome-wide significance thresholds are dependent on variant annotation increases the power to detect associations[3].

When the trait of interest is a quantitative trait, the trait is usually normalized and adjusted for confounding variables before performing the association. A linear regression model is then used to test for association between the trait and sequence variants, which may be written as

$$\mathrm{y_i} \backsim \alpha + \beta g_i + \varepsilon \ (*)$$

where i indexed the individual, $\mathrm{y_i}$ is the normalized traits and g_i is the genotype.

When the trait of interest is a binary trait, for instance a disease status, a logistic regression is used to test for association between the trait and sequence variants. The disease status is treated as the response variable and genotype as a covariate. In order to adjust for variables that correlate with disease status, such as age and sex, they are included in the model as nuisance variables.

2.2.1 Inheritance models

The most commonly used model in GWAS is the additive model (sometimes also called the multiplicative model for binary traits). The additive model assumes that the association depends additively on the minor allele, i.e. that the mean of a trait increases or decreases by β for each extra copy of the minor allele. To assume an additive model we can parameterize the genotype covariate in the regression model (*) as the allele count; 0 for non-carriers of the minor allele, 1 for heterozygous carriers and 2 for homozygous carriers. Then, for quantitative traits we are assuming that the mean of the trait changes linearly with the allele count and similarly for binary traits, that the risk on the log scale (log(OR)) changes linearly with the allele count.

However, traits and diseases can be inherited in different ways, for instance under a mode of recessive or dominant inheritance. The recessive model assumes that having two copies of the minor allele has an effect on the trait, while the dominant model assumes that having either one or two copies of the minor allele has an effect on the trait. Another form of inheritance is parent-of-origin mode of inheritance. The maternal model assumes that only the maternally inherited allele has an effect and conversely the paternal model assumes that only the paternally inherited allele has an effect.

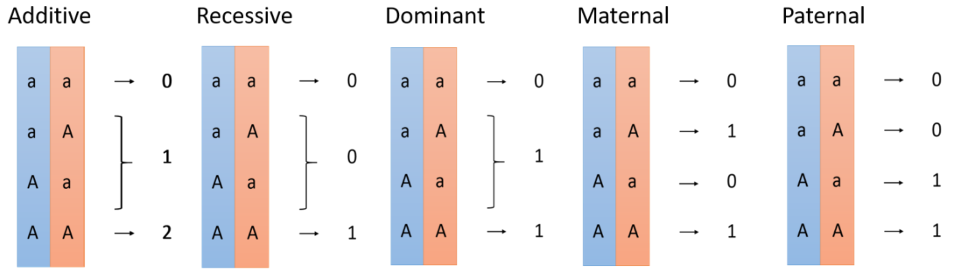

Figure 1. *Coding of genotypes under different models. The blue color represents alleles on the paternally inherited chromosome, and orange on the maternally inherited chromosome.* A *represents the minor allele and* a *the common allele.*

To implement tests for these models, we code the genotypes in different ways (Figure 1). When information is incomplete, the true genotypes of individuals are not known and we instead have allele probabilities ranging from 0 to 1. Assuming that the parental origin of the alleles can be determined with high accuracy, e.g. using long-range phasing and genealogy information[4], we let g_{mi} and g_{pi} be the allele probabilities for the maternally and paternally inherited alleles, respectively. Then the genotype used as a covariate in the additive model is coded as the sum of the probabilities; $g_{mi} + g_{pi}$. For the recessive model, the genotype is coded as the probability of individuals being homozygous carriers, i.e. the product of the allele probabilities, $g_{mi}g_{pi}$. When testing for maternal transmission, the genotype is defined as g_{mi} and when testing for paternal transmission, the genotype is defined as g_{pi}.

Sometimes, variants associating with phenotypes do not follow exactly the additive, recessive, dominant or parental model. For instance, a variant can have a small effect β on heterozygous carriers and a large effect γ on homozygous carriers, where γ is greater than 2β, as is assumed by the additive model. A full genotypic model can be used to assess the effect for heterozygous and homozygous carriers separately. In the full genotypic model we use two genotype covariates in the association model, g_{het} which is 1 for heterozygotes and 0 otherwise, and g_{hom} which is 1 for homozygotes and 0 otherwise.

2.2.2 Genetic effect on the variance of traits

For quantitative traits, GWAS usually aims to detect sequence variants that influence the mean of the trait. However, the variability of the trait can also be under genetic control, i.e. the variance of the trait can differ between different genotype groups. A sequence variant that associates with the mean of the trait can also affect the variance (Figure 2.a) and analyzing the effect of variants on the variance of traits can therefore improve our understanding of previously known variants. Furthermore, variants that do not affect the mean can affect the variance and searching for variants affecting the variance can potentially lead to the discovery of novel genotype-phenotype associations (Figure 2.b).

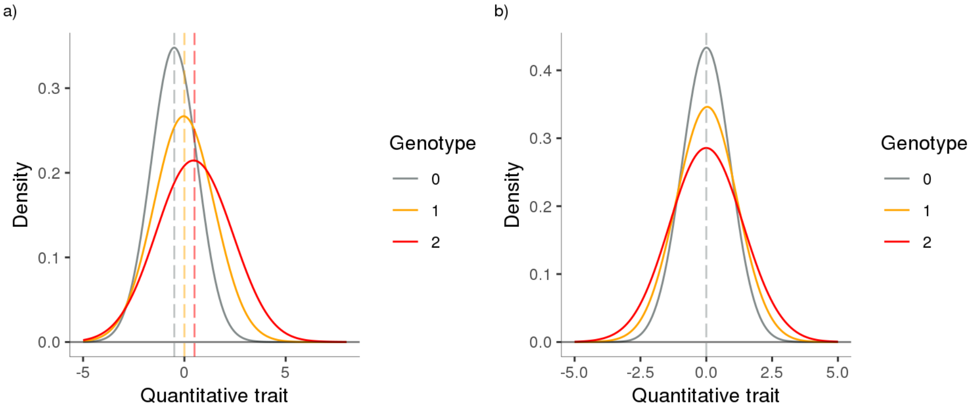

Figure 2. *Examples of sequence variants affecting the variance of a quantitative trait. The grey solid line shows the density of a trait for non-carriers and the grey dotted line shows the mean. The orange and red lines show the same for heterozygotes and homozygotes carriers, respectively. The examples are for **a)** a variant that associates with both increased mean and variance of a quantitative trait and **b)** a variant that associates with increased variance but does not affect the mean.*

While thousands of loci affecting the mean of complex traits have been discovered, analyzing the genetic effect on phenotypic variance is still rare and little is known about how commonly loci affect the variance of traits. However, increasing interest on the subject has led to identification of variance loci for several human traits, including the major histocompatibility complex (MHC) region for rheumatoid arthritis[5], *FTO* for body mass index[6], *SLC2A9* for serum urate[7], *LEPR* for C-reactive protein and *ICAM1* for soluble ICAM-1 levels[8]. More recently, few studies have reported loci affecting phenotypic variance, using the extensive UK Biobank data[9,10].

Methodologically, detecting sequence variants affecting phenotypic variance is performed with statistical tests for heteroscedasticity or variance heterogeneity. A distribution is said to be heteroscedastic if its variance is dependent on some variable, for example if the variance is unequal between some groups. A standard statistical test for heteroscedasticity is the Levene's test[11]. The Levene's test is a non-parametric F-test which compares the variable $z_{ij} = |y_{ij} - \bar{y}_i|$ between groups $i \in (1, \dots, k)$, where y_{ij} is the trait value for individual j in group i and $\bar{y}_i$ is the mean of the trait for group i. The test statistic is

$$W = \frac{(N-k)\sum_{i=1}^{k} n_i(\bar{z}_{i.} - \bar{z}_{..})^2}{(k-1)\sum_{k-1}^{n_i}\left(z_{ij} - \bar{z}_{i.}\right)^2},$$

where N is the total sample size and n_i is the sample size for group i, $\bar{z}_{i.}$ is the mean of all z_{ij} in group i and $\bar{z}_{..}$ is the mean of all z_{ij}. Under the null hypothesis of equal variance across the k groups, W follows an F distribution with $k-1$ and $N-k$ degrees of freedom. Several studies have used Levene's test, or the similar Brown-Forsythe test (replacing $\bar{y}_i$ with median as opposed to the mean), to test for genetic effect on the variance of traits[8,10,12]. Other statistical tests have been proposed to detect genetic effect on variance, such as the two stage double generalized linear model (DGLM)[13], and tests that allow for estimation of mean and variance effects jointly with a likelihood ratio test[14] and linear mixed models[9].

Previous studies have not taken into account that when multiple measurements are available per person, the total phenotypic variance is both due to between-subject (BS) variance and within-subject (WS) variance. In this thesis we suggest using variance models that estimate the effect of variants on the BS and WS variance by using a likelihood-ratio test and assuming the variance changes multiplicatively with the genotype.

2.3 Genetic epidemiology

2.3.1 Heritability

Heritability of a phenotype is the fraction of phenotypic variation in a population that is due to genetic inheritance. If we assume that a quantitative phenotype *y* may be partitioned into a genetic component, *G*, an environmental component, *E*, and a random noise, ε, such that

$$y = G + E + \varepsilon,$$

then

$$Var(y) = Var(G) + Var(E) + 2Cov(G, E) + Var(\varepsilon).$$

If the covariance between the genetic and environmental component is zero; $Cov(G, E) = 0$, then the broad-sense heritability of the phenotype y is defined as:

$$H^2 = \frac{Var(G)}{Var(y)}.$$

If we let A be the additive genetic component of y, then the narrow-sense heritability is defined as:

$$h^2 = \frac{Var(A)}{Var(y)}.$$

Most heritability estimates are based on comparing phenotypic similarities, or correlations, between relative pairs to the genetic sharing between relatives. Most commonly, twin studies are used to estimate heritability. For instance, Falconer‘s formula uses twice the difference between the phenotypic correlation of monozygotic and dizygotic twins to estimate broad-sense heritability[15]:

$$\widehat{H}^2 = 2(Cor_{MZ} - Cor_{DZ}).$$

A sequence variant with allele count g, allele frequency $Eg = f$, and mean effect β, $Ey|g = (g - f)\beta$, will create narrow-sense heritability that amounts to

$$Var(\beta g) = \beta^2 \mathrm{Var(g)} = 2f(1-f)\beta^2.$$

However, it is not clear what effect a sequence variant that associates with phenotypic variance has on heritability. In this thesis we explore how genetic effect on variance affect heritability estimates.

2.3.2 Mendelian randomization methods

Genetics can be a useful tool to explore the causal relationship between correlated phenotypes, using a method called Mendelian randomization (MR)[16]. Identifying modifiable causes to diseases are an important focus in many epidemiological studies. The existence of confounding variables that affect both the disease and the risk factor being explored can cause a spurious association. This can be problematic in observational epidemiological studies where adjusting for all confounding variables is usually impossible.

In MR methods, genetic variants are used as instruments, i.e. variables that associate with the risk factor of interest, are independent of possible confounders and associate with the disease outcome only through that risk factor. Genetic variants are often appropriate instrumental variables because genotypes are randomly assigned at birth and are not affected by environmental risk factors and are therefore independent of possible confounders.

In the simplest form, a single genetic variant can be used as an instrumental variable. When several sequence variants affect the risk factor of interest, one type of a MR method is to fit a regression model with their effect on the risk factor against their effect on the disease or outcome. A correlation between the effects suggests a causal relationship between the two traits. Another way is to aggregate the risk of the variants together into a genetic risk score (GRS) and use the GRS as an instrumental variable.

2.4 Phenotypes

A vast amount of phenotypic data is available in the deCODE database, that have been obtained from various sources. This section provides background information about the phenotypes used in this thesis.

2.4.1 Glucose levels and type 2 diabetes

Type 2 diabetes (T2D) is the form of diabetes that is characterized by high glucose levels in blood and insulin resistance. At the time Paper I was written, GWAS on T2D and glycemic traits had found 54 sequence variants associating with T2D and 36 variants associating with fasting glucose levels[17]. There is an overlap of 22 variants in these sets of associating variants, but surprisingly, the effect of the glucose variants on glucose does not predict their effect on T2D[17]. In particular, some variants strongly associate with glucose levels without increasing the risk of T2D.

2.4.2 The cornea

The cornea is the clear layer that covers the front part of the eye and corneal diseases are among the most common causes of visual loss[18]. The corneal endothelium is composed of a single layer of cells at the inner surface of the cornea (Figure 3). Its role is to maintain balance of fluids within the cornea and through a pump function it moves excess water and ions to and from the stroma[19]. The endothelial cells are hexagonally shaped and changes in the cell structure can lead to endothelial dysfunction, which is the most common cause for corneal transplantation[20]. The density of the cells is important and a minimum of 400-500 endothelial cells per square millimeter, is needed for the endothelium to function properly[21].

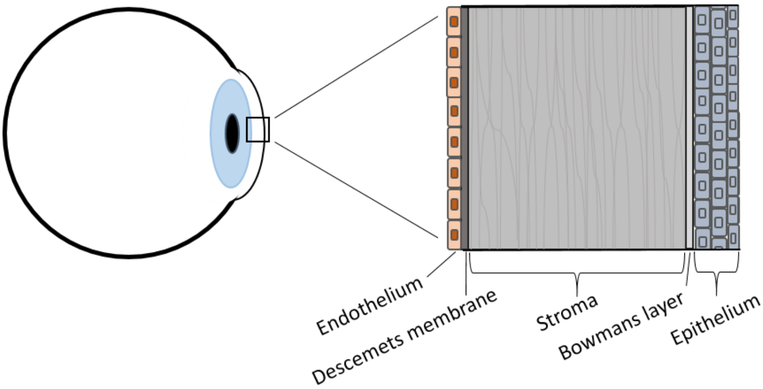

Figure 3. *A schematic figure of the cornea.*

A specular microscopy performs a non-invasive analysis of the corneal endothelium by capturing an image of the endothelial cell layer and provides measures of its structure including cell density (cells/mm^2), coefficient of cell size variation (CV), percentage of hexagonally shaped cells (HEX) and central corneal thickness (CCT). Examples of images of the corneal endothelium provided by the specular microscopy are shown in Figure 4. A healthy endothelium has high cell density and evenly sized and shaped cells (Figure 4.a). The example in Figure 4.b shows an endothelium with normal cell density, but high CV and low HEX since the cells vary more in size and shape. The example in Figure 4.c shows an endothelium with more evenly sized cells but extremely low cell density.

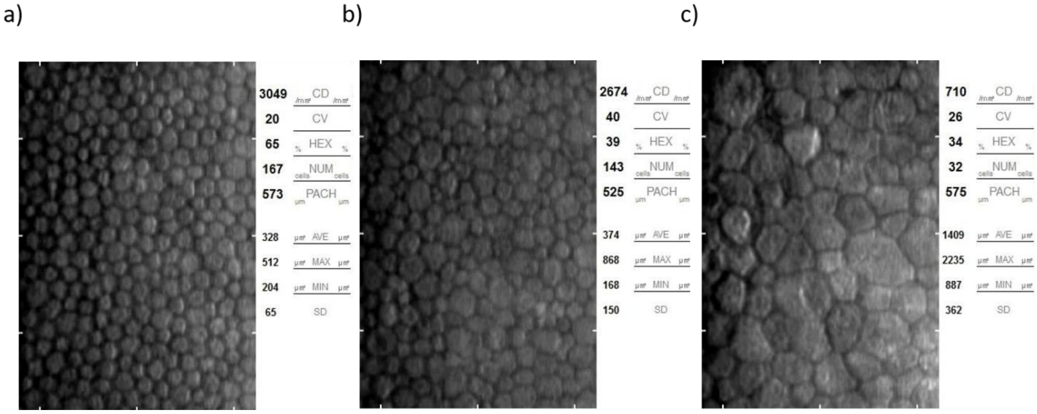

Figure 4. *Images of the corneal endothelium obtained with a specular microscopy.*

While cell density, HEX and CV are structural measures of the corneal endothelium, CCT is a measure of the thickness of all 5 layers of the cornea. Out of these four traits, CCT is the only one that has been explored in GWAS before, and then it was measured using other equipment[22–29]. Cell density, CV and HEX, are used in clinic as an indicator of the health of the cornea and to diagnose corneal endothelial diseases. Example of corneal diseases that are known to associate with the structure of the endothelial cells are Fuchs endothelial corneal dystrophy (FECD) and macular corneal dystrophy (MCD). It has also been observed that cell density can be reduced in patients with glaucoma[30]. However, our understanding of how these structural measures of the corneal endothelium relate to these diseases is still limited.

2.4.3 Hearing

Hearing loss is often categorized by age of onset. Prelingual hearing loss is present in 1-2 for every 1000 infants[31] and over 100 genes have been linked with prelingual or childhood-onset nonsyndromic hearing loss, 75% of which are inherited in a recessive manner (Hereditary Hearing Loss homepage). The most common type of hearing loss is age-related hearing impairment (ARHI) defined as the gradual decline of auditory function with age. It is usually caused by damage to the hair cells in the inner ears' cochlea, which are specialized receptors that receive sound waves and convert them into nerve signals that are transmitted to the brain by the auditory nerve[32]. Less is known about the genetics of ARHI and up until recently, only a few loci had been identified at the genome-wide significant level[33–38]. However, a recent study found 44 loci associating with self-reported hearing impairment obtained from UK Biobank[39].

The most common test used to measure auditory function is a pure tone audiometric test. Then an audiometer delivers pure tones at different frequencies or pitch measured in Hertz (Hz) and different intensity or loudness measured in decibels hearing levels (dB HL). This results in hearing thresholds per frequency for 0.5, 1, 2, 4, 6 and 8 kHz, which is the lowest intensity levels a subject hears the sound and a normal hearing threshold is ≤25 dB HL. ARHI is more common in men than women and is more common at the higher frequencies 4-8 kHz than at the lower frequencies 0.5-2 kHz. Pure tone average is defined as the average hearing threshold at the frequencies 0.5, 1, 2 and 4 kHz and represent the speech range.

Those who have ARHI are at more risk of tinnitus, the perception of a phantom sound often described as ringing or buzzing[40]. Despite the fact that 1-3% of the general population experience incessant tinnitus that severely affects their lives, treatments for tinnitus are lacking[41]. The heritability of tinnitus has been estimated to be 56% in a twin study[42] but several GWAS have not found variants associating with tinnitus[43] and the shared genetic causes between ARHI and tinnitus have not been broadly explored.

3 Aim

The main goals of this project are:

(i) Estimating the effect of known glucose variants on the BS and WS variance of glucose levels and examine the relationship between their effect on glucose levels and their effect on T2D. Furthermore, we explored how the effect of these variants on variance affect heritability estimates.

(ii) Using different statistical models to search for sequence variants associating with sensory measurements (auditory and ocular) obtained in the deCODE health study and explore the causal relationship between correlated traits and diseases.

4 Materials and Methods

The following chapter described the methods used for carrying out the aims of this thesis listed in Chapter 3. The appended papers provide further description of all methodologies used.

4.1 Study Subjects

4.1.1 The deCODE health study

The deCODE health study (DHS) is an ongoing population-based study in Iceland that is designed to improve our understanding of rare LOF mutations and other potentially high impact mutations. The participants in the study are a mixture of volunteers and carriers of rare predicted high impact mutations. Recruitment started in 2016 and in January 2020, the study included 11,484 Icelanders. The subjects were between 18 and 97 years of age at time of recruitment (43.6% men, mean age = 55.4, standard deviation (SD) = 14.5). Participants in the DHS go through a 3-hour baseline visit which includes verbal interviews about health and lifestyle, blood sample collection and several physical measurements. Additionally, they answer an online questionnaire and give permission to access health-related information including hospital data. A part of the collected measurements are various ocular measures and an air conduction audiometric test to measure hearing.

4.1.2 Paper I

The Icelandic study subjects in Paper I were 117,548 chip-typed individuals with glucose measurements obtained from three different laboratories; the National University Hospital of Iceland, Akureyri Hospital and Mjodd Laboratory. The Iranian study subjects were 10,437 chip-typed individuals with fasting glucose measurements obtained as a part of the Tehran Lipid and Glucose Study[44].

4.1.3 Paper II

The primary study subjects in Paper II were 6,266 Icelanders with endothelial images from a specular microscopy and measures of ocular biomechanics using an ocular response analyzer obtained from the DHS. When testing variants for association with ocular diseases, the glaucoma and corneal dystrophy cases were based on six ICD diagnostic codes obtained from Iceland and the UK Biobank study[45].

4.1.4 Paper III

The Icelandic study subjects in Paper III were from two non-overlapping datasets. Firstly, 11,484 individuals had air conduction audiometric measures obtained from the DHS where

4,140 were defined as cases (PTA>25 dB HL) and 7,344 as controls. Secondly, air conduction audiometric measures were obtained for 22,212 individuals from the National Institute of Hearing and Speech in Iceland (NIHSI). Around 44% of the measures were performed on children and since individuals with hearing problems are referred to the NIHSI, the dataset is highly skewed towards those with hearing impairment. We therefore defined, 9,619 individuals as AHRI cases (PTA>25 dB HL) and selected 298,609 population controls with no available hearing data, excluding all subjects that had participated in the DHS. The study subjects from the UK Biobank, were 108,175 individuals that reported hearing difficulty and 285,746 controls.

4.1.5 Ethical statement

The studies were approved by the Icelandic Data Protection Authority and the National Bioethics Committee (VSN-18-186 and VSNb2015120006/03.01 with amendments). Personal identifiers of the subjects were encrypted by a third-party system which is monitored by the Data Protection Authority. UK Biobank's scientific protocol and operational procedures were reviewed and approved by the North West Research Ethics Committee (REC Reference Number: 06/MRE08/65). The Tehran Lipid and Glucose Study has been approved by the National Research Council of the Islamic Republic of Iran (No. 121) and the Human Research Review Committee of the Endocrine Research Center, Shahid Beheshti University (M. C). All participating subjects, from Iceland, UK and Iran, provided written consent.

4.2 Genotypic data

The DNA samples from Iceland used in this thesis have been collected by deCODE through various studies since 1996.

4.2.1 Iceland

The number of WGS individuals at deCODE increases rapidly. At the time when Paper III was written, 49,708 Icelanders had been WGS and 34 million high quality sequence variants had been identified and imputed into 166,281 chip-typed individuals as well as relatives of the chip-typed. The process for WGS and the subsequent imputation has been described in previous reports[46–48].

4.2.2 UK Biobank

Two available sets of genotypes form the UK Biobank were used in this study. The first dataset consists of 26.5 million high quality variants from the Haplotype Reference Consortium (HRC) reference panel, imputed into chip-typed individuals of European ancestry[49]. The second dataset consists of 922 thousand variants which had been identified through whole exome sequencing (WES) of 49,960 study participants[50] and imputed into chip-typed individuals of European ancestry.

4.3 Association testing

4.3.1 Binary traits

A logistic regression is used to test for association between sequence variants and binary traits such as disease status. The trait is treated as the response variable and allele count as a covariate. To adjust for variables that correlate with the binary trait, such as age and sex, they are included in the model as nuisance variables. Let y_i be the disease status for individual $i \in \{1, 2,\ldots,n\}$, 1 if the individual has the disease and 0 otherwise, and g_i be the allele count. To test for an association between a sequence variant and the trait we assume that

$$L_i(\alpha,\beta,\gamma) = P(y_i|g_i,x_i)$$

$$logit(P(y_i|g_i,x_i)) \backsim \alpha + \beta g_i + \gamma^T x_i \ ,$$

where x_i are the nuisance variables and α, β and γ are regression coefficients. We then use a likelihood ratio test where the null hypothesis assumes no effect, $H_0\colon \beta = 0$, and use the asymptotic assumption that the likelihood ratio test statistic follows a χ^2 with one degree of freedom:

$$-2\log\left(\frac{max_{\alpha,\beta,\gamma}\, L(\alpha,\beta,\gamma)}{max_{\alpha,\gamma}\, L(\alpha,0,\gamma)}\right) \sim \chi_1^2.$$

The estimated odds ratio (OR) is then $\exp(\hat{\beta})$ where $\hat{\beta}$ is the maximum likelihood estimator of β. In our analysis, the allele counts are either from genotyping or integrates over all possible genotypes based on the phased imputations[1].

4.3.2 Quantitative traits

A generalized form of linear regression is used to test for association between sequence variants and quantitative traits. Before performing the association, the trait is adjusted for confounding variables and normalized. In particular, separately for each sex we adjust for age and other relevant variables using a generalized additive model[51] and then use rank-based inverse normal transformation to standardize the residuals. If an individual has multiple measures, the average of all measures, after standardization, is used. Let y be a vector of the normalized quantitative trait and g be a vector of the genotypes for the variant being tested. Then we assume that $\boldsymbol{y}$ is normally distributed with a mean that depend linearly on the genotype and a variance covariance matrix proportional to the kinship matrix:

$$y \ \backsim N\big(\alpha + \beta g,\ 2\sigma^2 \Phi\big),$$

where

$$\Phi_{ij} = \begin{cases} 1/2, & i = j \\ 2k_{ij}, & i \neq j \end{cases}$$

is the kinship matrix estimated from the Icelandic genealogy[46]. Then we use a likelihood ratio test to test for association and the maximum likelihood estimator of β is the estimated effect of the sequence variant on the trait. Obtaining the maximum likelihood estimators involve inverting the kinship matrix which is computationally intensive. We therefore split the individuals into smaller clusters for the calculations as described previously[46].

4.3.3 Variance model

The variance of the measurements of a quantitative trait y can be partitioned into between subject (BS) variance, within subject (WS) variance and residual variance.
Let N be the total number of subjects and for each subject $i \in \{1, \dots, N\}$ we have v_i measurments. We let $g_i \in (0,1,2)$ be the allele count of the variant being tested for individual i, and n_g be the number of individuals with genotype g. Testing for genetic effect on the BS variance we assume the following model

$$\bar{y}_i = N(\mu_{g_i}, \alpha^{g_i}\sigma^2), \qquad \text{(model 1)}$$

where $\bar{y}_i = \frac{1}{v_i}\sum_{j=1}^{v_i} y_{ij}$ is the average trait for individual i, μ_{g_i} is a nuisance parameter, since the mean is not of interest here, and α is the genetic effect on the BS variance. Under this model, the BS variance changes with the genotype multiplicatively in the following manner:

$$Var(\bar{y}_i|g_i = 0) = \sigma^2,$$
$$Var(\bar{y}_i|g_i = 1) = \sigma^2\alpha,$$
$$Var(\bar{y}_i|g_i = 2) = \sigma^2\alpha^2.$$

We then use a likelihood ratio test to test for genetic effect on the BS variance where the null hypothesis assumes no variance effect, H_0: $\alpha = 1$, and the alternative is H_1: $\alpha \neq 1$.
The log likelihood function of α and σ^2 is

$$l(\alpha, \sigma^2) \propto -\frac{1}{2}\left[(n_0 + n_1 + n_2)\log(\sigma^2) + (n_1 + 2n_2)\log(\alpha) + \frac{1}{\sigma^2}\left(R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}\right)\right]$$

where

$$R_g = \sum_{i:g_i=g} \left(\bar{y}_i - \bar{y}^*{}_g\right)^2, \qquad n_g = \sum_{i=1}^{N} I(g_i = g), \qquad \bar{y}^*{}_g = \frac{1}{n_g}\sum_{i:g_i=g} \bar{y}_i.$$

To find the maximum likelihood estimators, $\widehat{\sigma^2}$ and $\hat{\alpha}$, we solve

$$\frac{\partial l}{\partial \sigma^2} = 0$$
$$\Leftrightarrow \; -\frac{1}{2}\left[\frac{(n_0 + n_1 + n_2)}{\sigma^2} - \left(R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}\right)\frac{1}{(\sigma^2)^2}\right] = 0$$
$$\Leftrightarrow \; (n_0 + n_1 + n_2) - \left(R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}\right)\frac{1}{\sigma^2} = 0$$

$$\Leftrightarrow \ \sigma^2 \ = \ \frac{R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}}{n + n_1 + n_2}$$

assuming that $\sigma^2 \neq 0$. And then we solve

$$\frac{\partial l(\alpha, \widehat{\sigma^2})}{\partial \alpha} = -\frac{1}{2}\left[(n_0 + n_1 + n_2)\log\left(\frac{R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}}{n + n_1 + n_2}\right) + (n_1 + 2n_2)\log(\alpha) \right.$$
$$\left. + (n_0 + n_1 + n_2)\right] = 0$$

$$\Leftrightarrow \ -\frac{1}{2}\left[\frac{(n_0 + n_1 + n_2)\cdot(-\frac{R_1}{\alpha^2} - 2\frac{R_2}{\alpha^3})}{R_0 + \frac{R_1}{\alpha} + \frac{R_2}{\alpha^2}} + \frac{(n_1 + 2n_2)}{\alpha}\right] = 0$$

$$\Leftrightarrow \ -\frac{\alpha^3}{2}[\alpha^2 R_0(n_1 + 2n_2) + \alpha R_1(n_2 - n_0) - R_2(n_1 + 2n_0)] = 0$$

$$\Leftrightarrow \ \alpha = \frac{-B + \sqrt{B^2 - 4AC}}{2},$$

where $A = R_0(n_1 + 2n_2),\ B = R_1(n_2 - n_0),\ C = -R_2(n_1 + 2n)$, assuming $\alpha > 0$. We note that $-4AC = 4R_0(n_1 + 2n_2)R_2(n_1 + 2n)$ is always positive and since $n_0 > n_2$ we have that $B > 0$. Therefore, we have that $\sqrt{B^2 - 4AC} > B \ \Rightarrow \ \frac{-B+\sqrt{B^2-4AC}}{2} > 0$ and $\frac{-B-\sqrt{B^2-4AC}}{2} < 0$.

Then the maximum likelihood estimators are:

$$\widehat{\sigma^2} = \frac{R_0 + \frac{R_1}{\hat{\alpha}} + \frac{R_2}{\hat{\alpha}^2}}{v_0 + v_1 + v_2}$$
$$\hat{\alpha} = \frac{-B + \sqrt{B^2 - 4AC}}{2}.$$

The test statistic is $D = 2(l(\hat{\alpha}, \widehat{\sigma^2}) - l(1, \widehat{\sigma^2}_{\alpha=1}))$ and the asymptotic distribution of D is χ_1^2.

Similarly, we test for a WS variance effect by fitting a model where the WS variance is allowed to change multiplicatively with the genotype.
Let N be the total number of subjects and for each subject $i \in \{1, \ldots, N\}$ we have v_i measurments. We let $g_i \in (0,1,2)$ be the allele count of the variant being tested for individual i, and n_g be the number of individuals with genotype g. We assume that

$$y_{ij} = \ N(\mu_{g_i}, \alpha^{g_i}\sigma^2), \qquad \text{(model 2)}$$

where y_{ij} is the jth measure of the trait for individuals i, and use a likelihood ratio test to test for genetic effect on the WS variance where the null hypothesis is H_0: $\alpha = 1$ and the alternative is H_1: $\alpha \neq 1$.
The log likelihood function of α and σ^2 is

$$l(\alpha, \sigma^2) = -\frac{1}{2}\Big[(n_0 + n_1 + n_2)\log(\sigma^2) + (n_1 + 2n_2)\log(\alpha) + \frac{1}{\sigma^2}\Big(RSS_0 + \frac{RSS_1}{\alpha} + \frac{RSS_2}{\alpha^2}\Big)\Big]$$

where,

$$RSS_g = \sum_{i:g_i=g}\sum_{j=1}^{v_i}\left(y_{ij} - \bar{y}_i\right)^2 , \qquad \bar{y}_i = \frac{1}{v_i}\sum_{j=1}^{v_i} y_{ij} , \qquad n_k = \sum_{i:g_i=k} v_i - 1.$$

We can then derive the maximum likelihood estimators in the same way as for model 1:

$$\hat{\sigma} = \frac{RSS_0 + \frac{RSS_1}{\alpha} + \frac{RSS_2}{\alpha^2}}{n_0 + n_1 + n_2}$$

$$\hat{\alpha} = \frac{-B + \sqrt{B^2 - 4AC}}{2},$$

where $A = RSS_0(n_1 + 2n_2)$, $B = RSS_1(n_2 - n_0)$ and $C = -RSS_2(n_1 + 2n_0)$. The test statistic is $D = 2(l(\hat{\alpha}, \widehat{\sigma^2}) - l(1, \widehat{\sigma^2}_{\alpha=1}))$ and $D \sim \chi^2_1$.

The BS and WS variance models are sensitive to the distribution assumptions of the data. Before fitting the BS variance model, $\bar{y}_i = N(\mu_{g_i}, \alpha^{g_i}\sigma^2)$, then for $v = 1, \dots, t-1, \geq t$ we re-standardize the subset of individuals that have the same number of measurements to follow a standard normal distribution. Here t is a threshold, chosen to be $t = 15$ in the data including only fasting glucose levels and $t = 50$ in the data including all glucose levels. Before fitting the WS variance model, $y_{ij} = N(\mu_{g_i}, \alpha^{g_i}\sigma^2)$, then for $v = 1, \dots, t-1, \geq t$ we re-standardize the subset of residual sum of squares $\sum_{j=1}^{v_i}\left(y_{ij} - \bar{y}_i\right)^2$ where $v_i = v$ to fit a chi-squared distribution with $v - 1$ degrees of freedom.

We use simulations to examine the performance of the BS and WS variance models, in terms of the rate of true and false positive associations and compare it to the performance of the Levene's test.

For evaluating the BS variance test, we simulate the phenotype $\bar{y}_i$ for different sample size $N \in \{10000, \dots, 500000\}$, using the model $\bar{y}_i = N(0, \alpha^{g_i}\sigma^2)$, where σ^2 was 1 and the BS variance effect α ranged from 1 to 1.06 in increments of 0.01. The genotype g_i was simulated from a binomial distribution with the probability parameter as the MAF of 0.3. For each N and each α, 1000 independent phenotypes and genotypes were simulated. To investigate the false positive rate of the variance models when the phenotype is not normally distributed, we additionally performed phenotype simulation assuming the phenotype to have a χ^2-distribution with 1 and 10 degrees of freedom. Further, to investigate if variance effects show up artificially for variants with large mean effects, we simulate phenotypes assuming a normal distribution with mean effects ranging from 0.1 to 0.7 SD for different minor allele frequencies and test for BS variance effect.

When evaluating the WS variance test, for each i we first simulate υ_i, the number of measurements per person, and then for each i and $j \in \{2, .., \upsilon_i\}$ we simulate the phenotype using the model $y_{ij} = \delta_i + \varepsilon_{ij}$, where $\delta_i \sim N(0, \sigma_{BS}^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_{WS}^2)$. We assume no effect on the BS variance and let $\sigma_{BS}^2 = 1$, while we assume genetic effect on the WS variance and that the WS variance is smaller than the BS variance and let $\sigma_{WS}^2 = \alpha^{g_i} \cdot 0.2$. As for the BS variance model we perform the simulation for different sample size $N \in \{10000, \dots, 100000\}$ and for different WS variance effects $\alpha \in \{1, 1.01, \dots, 1.06\}$.

4.3.4 Meta-analysis

In Papers II and III, when we meta-analyze GWAS results from different cohorts, we use a fixed-effects inverse variance method[52]. Sequence variants from deCODE and the UK Biobank were matched on position and alleles and the method is based on standard errors and effect estimates from the GWAS.

4.3.5 Adjusting for population stratification

We use LD score regression[53] to account for distribution inflation in the datasets that can be due to relatedness and population stratification. The χ^2 statistics from the GWAS were regressed against the LD score for a set of 1.1 million variants. The intercept from the regression was used as correction factors.

4.4 Genetic effect on variance and heritability

How variants affecting the BS variance of traits affect heritability estimates is unclear. To investigate this, we assume the following model

$$y_i = \beta(g_i - f) + g_i' F_i + \varepsilon_i \sim N(0,1) \ (*)$$
$$F_i \sim N(0, {\sigma_F}^2)$$
$$\varepsilon_i \sim N(0, {\sigma_0}^2)$$

where y_i is the trait measure for individual i, g_i is the genotype or allele count, f is the allele frequency, β is the mean effect and $g_i' = 0,1,\sqrt{2}$ for each genotype 0,1 or 2. We assume that g_i, F_i, and ε_i, are all pairwise independent for each i and that different variables are independent between different individuals. However, all the variables may be correlated between individuals.

We note that under this model (*) the phenotypic variance partitions as follows:

$$\begin{aligned} 1 = Var(y_{ij}) &= Var(\beta(g_i - f)) + Var(g_i' F_i) + Var(\varepsilon_{ij}) \\ &= \beta^2 Var(g_i) + E(g_i'^2) Var(F_i) + Var(\varepsilon_{ij}) \\ &= 2f(1-f)\beta^2 + 2f{\sigma_F}^2 + {\sigma_0}^2. \end{aligned}$$

Under the (*) model, the BS variance increases linearly with the genotype:

$$Var(y_{ij}|g_i = 0) = \sigma_0{}^2$$
$$Var(y_{ij}|g_i = 1) = \sigma_0{}^2 + \sigma_F{}^2$$
$$Var(y_{ij}|g_i = 2) = \sigma_0{}^2 + 2\sigma_F{}^2$$

When estimating the BS variance effects before we used the model $y_{ij} \sim N(\mu_{g_i}, \alpha^{g_i}\sigma_0{}^2)$ (**) for convenience. When $\alpha \in [0.9; 1.1]$, then $(\alpha^2 - 1) \cong 2(\alpha - 1)$ and:

$$Var(y_{ij}|g_i = 0) = \sigma_0{}^2$$
$$Var(y_{ij}|g_i = 1) = \sigma_0{}^2\alpha = \sigma_0{}^2 + \sigma_0{}^2(\alpha - 1)$$
$$Var(y_{ij}|g_i = 2) = \sigma_0{}^2\alpha^2 = \sigma_0{}^2 + \sigma_0{}^2(\alpha^2 - 1) \cong \sigma_0{}^2 + \sigma_0{}^2 2(\alpha - 1)$$

so that these models, (*) and (**), are equivalent up to the first order around the null with

$$\sigma_F{}^2 \cong \sigma_0{}^2(\alpha - 1).$$

We will now show that the (*) model predicts that siblings sharing identical high variance genotypes have higher covariance than siblings that share identical low variance genotypes and that the difference between their covariances may be used to estimate the covariance between the *F*s of siblings.

Under the (*) model, the covariance between a pair of siblings *i* and *j*, given their genotypes, is

$$Cov_{sib}(y_i, y_j|g_i, g_j) = Cov_{sib}(\varepsilon_i, \varepsilon_j) + \sqrt{g_i g_j}Cov_{sib}(F_i, F_j).$$

In particular:

$$Cov_{sib}(y_i, y_j|g_i = g_j = 0) = Cov_{sib}(\varepsilon_i, \varepsilon_j)$$
$$Cov_{sib}(y_i, y_j|g_i = g_j = 1) = Cov_{sib}(\varepsilon_i, \varepsilon_j) + Cov_{sib}(F_i, F_j)$$
$$Cov_{sib}(y_i, y_j|g_i = g_j = 2) = Cov_{sib}(\varepsilon_i, \varepsilon_j) + 2Cov_{sib}(F_i, F_j).$$

Therefore, we used how the correlation between the phenotypes of siblings differs between genotype groups to estimate $Cov_{sib}(F_i, F_j)$ in the following way.

For *N* sibling pairs where both siblings in pair *i* are carrying genotype g_i and have phenotypes y_{1i} and y_{2i}, we calculated the mean phenotype for each genotype group *g*:

$$\bar{y}_g = \frac{1}{2N_g}\sum_{i:g_i=g} y_{1i} + y_{2i},$$

where N_g was the number of sibling pairs sharing genotype *g*. Then we defined

$$c_i = (y_{1i} - \bar{y}_{g_i})(y_{2i} - \bar{y}_{g_i}).$$

For each genotype g

$$C_g = \frac{1}{N_g} \sum_{i:g_i = g} c_i$$

is then an estimate of the covariance between siblings that share the genotype g, i.e. and estimate of $Cov_{sib}(y_i, y_j | g_i = g_j = g)$.
We then performed a weighted linear regression $C_g = \gamma g + \delta$ where we weight by $N_g / Var(y_{ij}|g)^2$. Then $\hat{\gamma}$ is an estimate of $Cov_{sib}(F_i, F_j)$.
To obtain a P-value, we estimate the covariance trend for all sequence variants in the genome, and for each variant we want to test, we compare its trend $\hat{\gamma}$ to the distribution of trends for all variants with similar frequency, i.e. all variants in the interval (f-0.025, f+0.025), where f is the frequency of the variant we are testing. This method may also be used to estimate the covariance between *F*s of other relative pairs such as parent offspring pairs.

The correlation between relative pairs is the ratio of their covariance and the geometric mean of their phenotypic variances. If we assume model (**) and that the covariance given the genotype changes multiplicatively,

$$Cov(y_i, y_j | g_i = g_j = g) = \gamma^g Cov(y_i, y_j | g_i = g_j = 0),$$

then the correlation given the genotype is:

$$Cor(y_i, y_j | g_i = g_j = g) = \frac{\gamma^g \, Cov(y_i, y_j | g_i = g_j = g)}{\alpha^g \, Var(y|g=0)} = \frac{Cov(y_i, y_j | g_i = g_j = g)}{Var(y|g=0)} \left(\frac{\gamma}{\alpha}\right)^g$$

Therefore, γ/α is the correlation trend. Therefore, if the covariance increase γ is larger than the variance increase α, the correlation is also increasing with the genotype.
Variants that effect the WS variance do not influence the covariance between subjects. But the total variance is affected, and therefore the correlation between subjects. If we assume $y_{ij} = \beta(g_i - f) + \varepsilon_{ij}$, where , $\varepsilon_{ij} \sim N(0, \alpha^{g_i}\sigma^2)$ and α is the WS variance effect then

$$Cor(y_{ij}, y_{kl} | g_i, g_k) = \frac{Cov(y_{ij}, y_{kl} | g_i, g_k)}{\sqrt{Var(y_{ij}|g_i) Var(y_{kl}|g_k)}} = \frac{Cov(y_{ij}, y_{kl})}{\alpha^{(g_i + g_k)/2} Var(y|g=0)}$$

So if the within subject variance increases with the genotype, and there is no genetic effect on the covariance, then the correlation decreases with the genotype.

4.5 Correlation between effect sizes

When assessing the relationship between the effect of genetic variants on two different traits, we use a simple weighted linear regression model, where each variant is weighted by f(1-f) where f is the MAF for that variant. In that way, rare variants have less weight in the regression model than common variants.

4.6 Genetic risk score

Genetic risk scores were constructed by combining the effect allele count for each variant, weighted by its effect. I.e. if we let m_{vi} and p_{vi} be the genotype probability for individual i and sequence variant v at the maternally and paternally inherited chromosomes, then the GRS, based on n variants, for individual *i* is

$$grs_i = \sum_{v=1}^{n} (m_{vi} + p_{vi})\beta_v,$$

where β_v is the genetic effect of variant *v*. For the variance GRS, the effect on the variance are used instead of the effects on the mean.

In Paper III, a genetic risk score is constructed for n variants associating under the additive model and m variants associating under the recessive model. As was described in section 2.2.1, the genotype used as a covariate in the additive model is coded as the sum of the probabilities; $m_{vi} + p_{vi}$ and for the recessive model, the genotype is coded as the probability of individuals being homozygous carriers, i.e. the product of the allele probabilities, $m_{vi} \times p_{vi}$. We therefore define the GRS for individual *i* as:

$$grs_i = \sum_{v=1}^{n} (m_{vi} + p_{vi})\beta_v + \sum_{v=1}^{m} (m_{vi} \times p_{vi})\gamma_v,$$

where β are the effects of the variants detected with the additive model and γ are the effects of the variants detected with the recessive model.

We note that this definition of the GRS relies on determination of parental origin, e.g. using long-range phasing and genealogy information[4]. If no parental information is available, one can define the GRS as

$$grs_i = \sum_{v=1}^{n} g_v\beta_v + \sum_{v=1}^{m} max(g_v - 1,0)\gamma_v,$$

where g_v is the allele count for sequence variant *v*.

5 Summary of key results

The following chapter summarizes the main findings of this project. The appended papers provide more detailed results.

5.1 Paper I - Sequence variants affecting the variance of glucose levels

5.1.1 Performance of the variance models

We performed simulations to estimate the power and false positive rates of the BS and WS variance models, as described in section 4.3.3.

Figure 5 shows the power of the BS variance and Levene's test, defined as the rate of associations with P<0.05 divided by the total number of tests across 1000 simulations. The power is shown for different sample sizes (x axis) and different true variance effect (shown in different colors). As the figure shows, the BS variance test is more powerful than the Levene's test. For each sample size, the null model with no variance effect was also simulated and the false positive rate of the BS variance test was on average 0.048 (0.006 SD), similarly to the Levene´s test; 0.049 (0.006 SD).

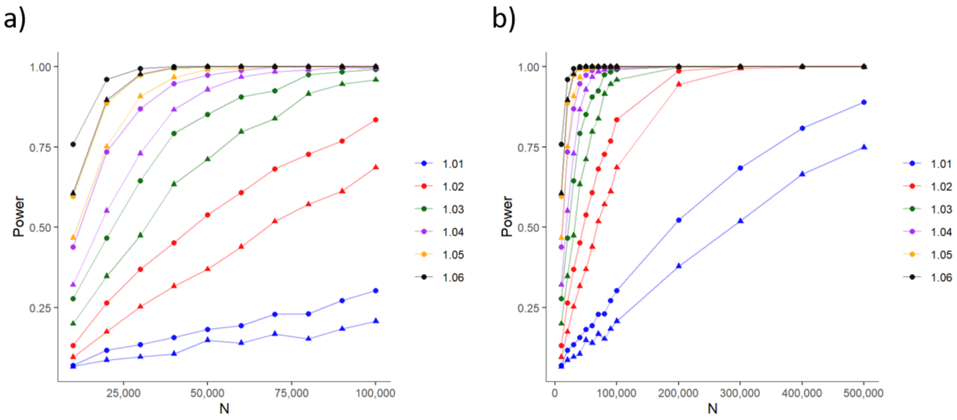

Figure 5. *The power of detecting BS variance effect for different simulation scenarios using the BS variance model (round dots) and the Levene's test (triangles) for different true variance effects (between 1.01 and 1.06) and different sample sizes. Figure* ***a)*** *shows the results for the simulated phenotype using sample sizes from 10,000 to 100,000 individuals and* ***b)*** *shows additionally the sample sizes from 200,000 to 500,000 individuals.*

To evaluate the false positive rate of the BS variance test when the phenotype does not follow a normal distribution, we additionally simulated the phenotypes using χ^2-distribution with 1 and 10 degrees of freedom. For this analysis we simulated phenotypes for 50,000 individuals. The BS variance test is sensitive to distribution assumptions and the false positive rate is high if no standardization is performed (Table 1). But after standardizing the phenotype, so that subsets of individuals with the same number of measurements follow a standard normal distribution, the false positive rate is well calibrated (Table 1).

Table 1. *The FPR of the BS variance test and Levene's test for different phenotype distributions estimated from 1000 simulations for 50,000 individuals.*

Phenotype distribution	**Levene's test**	**BS variance test**	**BS variance test after standardization**
χ_1^2	0.045	0.459	0.049
χ_{10}^2	0.051	0.113	0.039

To evaluate if large mean effects can cause artificial variance effects, we furthermore simulated phenotypes assuming a normal distribution with constant variance and mean effects ranging from 0.1 to 0.7 SD for different minor allele frequencies. We tested the simulated phenotypes for BS variance effect for 1000 simulations and computed the fraction of significant BS variance effects (Figure 6). For mean effects below 0.5 SD, the fraction of significant variance tests for 1000 simulations was around 5%. For mean effects larger than 0.5 SD, a variance effect artificially shows up alongside the mean effects. We note, that the vast majority of reported mean effects for common variants are below 0.5 SD.

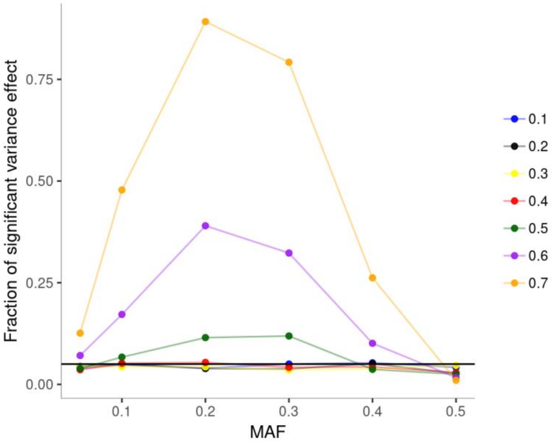

Figure 6. *The fraction of significant BS variance effects for a simulated normal phenotype for 100,000 individuals using different mean effects (between 0.1 and 0.7 SD) and different minor allele frequency (MAF) of the simulated variants. The fraction was computed as the number of BS variance associations with P<0.05 divided by the total number of tests across 1000 simulations for each simulation scenario.*

Testing the WS variance model, we first simulate the number of measurements per person (Figure 7.a) and then perform 1000 simulations of the phenotype assuming genetic effect on the WS variance ranging from 1.01 to 1.06 (Figure 7.b). The false positive rate for the WS variance model was 0.049 for 1000 simulations under the null model. We note that the Levene's test does not capture the WS variance effects (Figure 7.b).

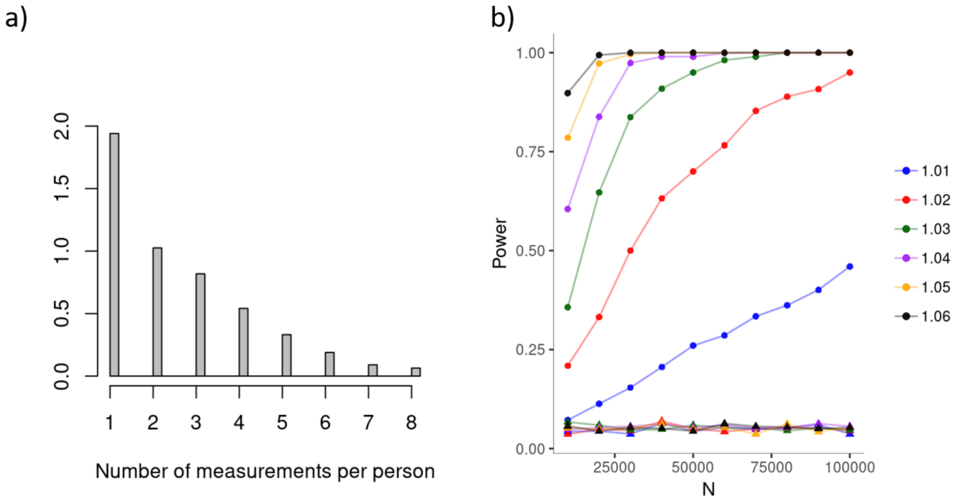

Figure 7. a) *The distribution of the simulated number of measurements per person.* ***b)*** *The power of detecting WS variance effect for different simulation scenarios using the WS variance model (round dots) and the Levene's test (triangles) for different true WS variance effects (between 1.01 and 1.06) and different sample sizes.*

5.1.2 Data summary

In total, 941,087 glucose measures were available for 117,548 individuals, out of which 69,142 individuals had fasting glucose measures. The average number of measurements per person was 8 for all glucose measures (Figure 8) and 3 for fasting glucose measures.

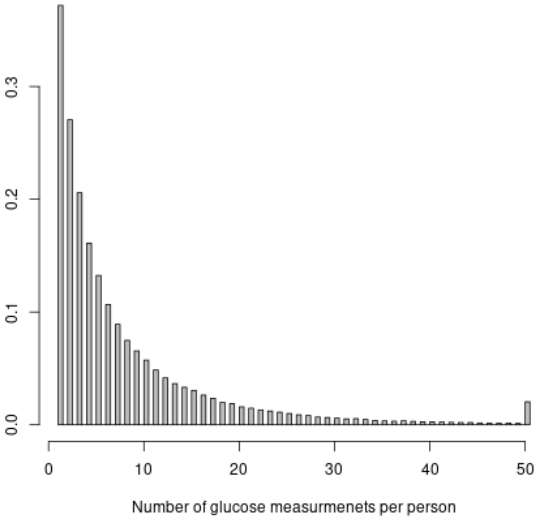

Figure 8. *A histogram that shows the number of all glucose measurements per person.*

Figure 9 shows how glucose levels are higher for men than women and that both the mean and variance of glucose levels increases with age. Of the subjects, 7.5% had T2D diagnosis or were on diabetes medication[54] and 0.3% had type 1 diabetes (T1D). All data analysis was performed on the subjects with fasting glucose measurements (dataset I) and for secondary analysis we used individuals with any glucose measurements (dataset II).

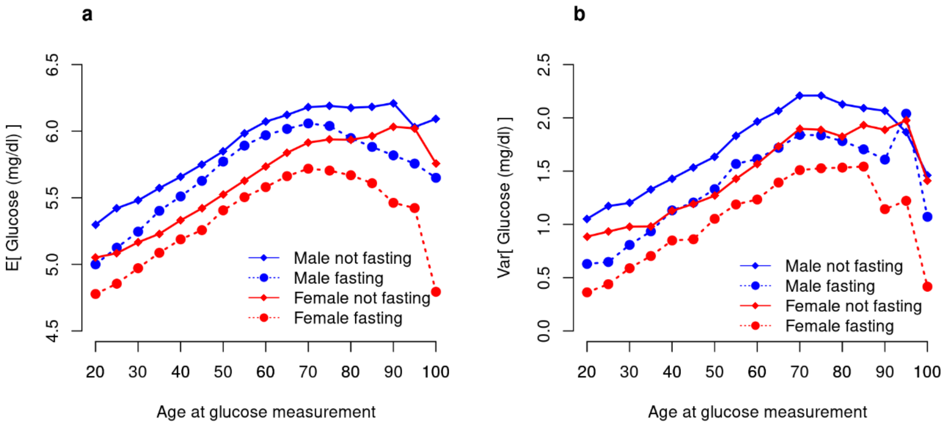

Figure 9*. Mean and variance of glucose levels against age.* ***(a****) The average and* ***(b)*** *the variance of fasting and not fasting glucose (mmol/L) levels per age group for both sexes.*

5.1.3 Effect of glucose variants on variance in glucose levels

For the 36 variants that have been previously reported to associate with the mean of fasting glucose levels[17], we estimated their effect on the BS variance of fasting glucose levels. Three of the variants associated with BS variance (Table 2). To investigate whether their effect on BS variance was due to diabetic individuals and medication intake, we estimated the effects using datasets I and II after removing individuals with T2D and T1D and the effects remained significant. Interestingly, among the sequence variants that increase glucose levels on average, some associate with increased BS variance and others with decreased BS variance. The variant in *TCF7L2* is the strongest common T2D associating variant.

Table 2. *The association of three glucose variants on the BS variance of glucose levels.*

						Diabetic subjects removed			
		Glucose		**Fasting Glucose**		**Glucose**		**Fasting Glucose**	
Gene	**MAF**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**
G6PC2	0.30	-0.075	2.0×10^{-28}	-0.076	7.5×10^{-18}	-0.048	3.5×10^{-12}	-0.060	2.2×10^{-11}
GCK	0.14	-0.075	1.8×10^{-16}	-0.078	4.9×10^{-11}	-0.058	1.6×10^{-10}	-0.059	6.4×10^{-7}
TCF7L2	0.30	0.052	2.5×10^{-14}	0.055	4.5×10^{-10}	0.036	1.7×10^{-7}	0.041	5.2×10^{-6}

We also tested the 36 glucose variants for association with the WS variance of fasting glucose levels. Three variants associated significantly with decreased WS variance (Table 3) in all datasets. Two of them, variants in *GCK* and *G6PC2*, are the same variants that associated with BS variance.

Table 3. *The association of three glucose variants on the WS variance of glucose levels.*

						Diabetic subjects removed			
		Glucose		**Fasting Glucose**		**Glucose**		**Fasting Glucose**	
Gene	**MAF**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**
G6PC2	0.30	-0.040	7.9×10^{-50}	-0.048	1.0×10^{-13}	-0.041	1.6×10^{-42}	-0.049	4.1×10^{-12}
GRB10	0.34	-0.018	2.3×10^{-12}	-0.026	2.8×10^{-5}	-0.016	7.6×10^{-9}	-0.0273	8.0×10^{-5}
GCK	0.14	-0.020	4.1×10^{-8}	-0.029	5.2×10^{-4}	-0.022	1.7×10^{-8}	-0.032	7.0×10^{-4}

The fasting glucose mean effects of the 36 mean effect variants are only weakly correlated with their effects on variance (correlation with BS effect: r^2=0.11, P=0.026, correlation with WS effect: r^2=0.11, P=0.026), while the BS variance effects and the WS variance effects are highly correlated (r^2=0.47, P=2.4×10^{-6}).

When testing for variance effects genome wide, we did not observe any variants that associate with glucose variance but not mean glucose levels. For the fasting glucose data set, the correction factors were $\lambda = 1.14$ and $\lambda = 1.21$ when estimating between-subject and within-subject variance effects, respectively.

5.1.4 Replication of variance effects

To validate the observed variance effects, we analyzed fasting glucose measures from 10,574 Iranians with 1 to 4 available measurement per person. We replicated 5 out of 6 associations (Table 4).

Table 4. *The association of glucose variants on the BS and WS variance of glucose levels in the Icelandic and Iranian datasets.*

			Iceland				**Iran**	
			Glucose		**Fasting Glucose**		**Fasting Glucose**	
Variance model	**Gene**	**MAF**	**Effect**	**P-value**	**Effect**	**P-value**	**Effect**	**P-value**
BS	*G6PC2*	0.30	-0.075	2.0×10^{-28}	-0.076	7.5×10^{-18}	-0.111	7.0×10^{-6}
BS	*GCK*	0.14	-0.075	1.8×10^{-16}	-0.078	4.9×10^{-11}	-0.117	2.7×10^{-5}
BS	*TCF7L2*	0.30	0.052	2.5×10^{-14}	0.055	4.5×10^{-10}	0.095	8.3×10^{-6}
WS	*G6PC2*	0.30	-0.040	7.9×10^{-50}	-0.048	1.0×10^{-13}	-0.036	2.1×10^{-2}
WS	*GRB10*	0.34	-0.018	2.3×10^{-12}	-0.026	2.8×10^{-5}	-0.012	4.1×10^{-1}
WS	*GCK*	0.14	-0.020	4.1×10^{-8}	-0.029	5.2×10^{-4}	-0.073	2.4×10^{-5}

5.1.5 Effect on glucose levels vs. effect on T2D risk

Out of the 36 glucose variants, 22 also associate with T2D in a T2D meta-analysis of European ancestry with 12,171 cases and 56.862 controls. However, their effect on glucose does not predict their effect on T2D, where the effects are only weakly correlated (r^2=0.02, P=0.21). For instance, the variants in *GCK* and *G6PC1* that increase glucose levels substantially, do not have any effect on the risk of T2D.

Interestingly, when we look at the relationship between BS variance effect and T2D risk, we see that BS variance effect predicts the effect on T2D quite well (r^2=0.38, P=3.3×10^{-5}). By fitting a regression model where the T2D effect is regressed against the mean and BS variance effect of fasting glucose, we get r^2=0.61 (P for adding effect on BS variance =5.7×10^{-8}). The *TCF7L2* variant has the greatest positive BS variance effect and the greatest effect on the risk of T2D and as such carries substantial weight in the regression. However, when we remove it from the analysis, the mean and BS variance effects of the other variants still predict T2D risk (r^2=0.46, P=7.1×10^{-6} for adding the BS effect) and the remaining variants predict the T2D effect of the *TCF7L2* variant to be high, although not as high as the observed OR (observed OR = 1.33 (95% CI: 1.29-1.37), predicted OR = 1.17). Performing the regression of T2D effect against the mean and BS effect using all glucose measurements instead of only fasting glucose gives consistent prediction (r^2=0.70 and P=4.1×10^{-10}). These results show that sequence variants that increase both the mean and BS variance tend to increase the risk of T2D more than variants that increase the mean but reduce the variance.

The effect on WS variance is a worse predictor of T2D effect than the effect on BS variance and adding the effect on WS variance to the regression model does not improve the prediction (P=0.091 for adding effect on WS variance).

5.1.6 Possible interaction

A possible explanation to why a sequence variant could affect the BS variance of a trait is an interaction between the variant and some environmental factor. BMI measures was available for 39,986 individuals and to investigate a possible interaction between the glucose variants and BMI on fasting glucose, we fitted the following model for each of the 36 variants:

$$Glucose = \gamma_1 g + \gamma_2 BMI + \gamma_3 (g \times BMI) + \varepsilon.$$

Seven variants had nominally significant interactions effect γ_3 with BMI ($P < 0.05$), including the *TCF7L2* variant ($P = 3.6\times10^{-3}$) and for the 36 variants, the interaction effects correlated with the effects on BS variance (r^2=0.12, P=0.020). That suggests that the effect of these variants on glucose levels are affected by the environment, but only a small fraction of their effect on BS variance are explained by interaction with BMI.

5.1.7 Genetic risk score

We constructed mean and variance GRSs for the 36 glucose variants, by using the mean effects and BS variance effects as weights (see section 4.6). Both GRSs associated with T2D ($P<3.1\times10^{-39}$). Further, in joint analysis of the fasting glucose risk scores, adding the variance GRS to the mean GRS increased the residual Nagelkerke's[55] pseudo r^2 from 0.4% to 1.0% ($P=5.4\times10^{-67}$).

Figure 10 shows the percentage of T2D cases in each quintile of the GRSs and how the effect of variants on BS variance have an impact on genetic risk prediction of T2D comparable to their effect on the mean.

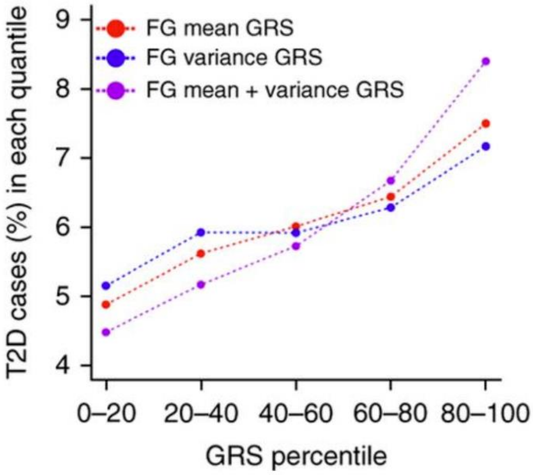

Figure 10. *Fasting glucose (FG) GRS based on the 36 fasting glucose variants.*

5.1.8 Heritability

Fasting glucose measures and genotype information was available for 35,965 sibling pairs and 38,527 parent-offspring pairs. For each of the 36 glucose variants, we computed the covariance for the pairs that shared the same genotype and estimated the trend between the genotype and the covariance. This was done for sibling pairs and parent-offspring pairs separately and then we computed the mean covariance trend. The mean covariance trend for the 36 variants was correlated with the BS variance effect (r^2=0.22, P=2.1×10^{-3}).

The variant with the strongest covariance trend was the variant in *TCF7L2* with 17.6% increased covariance per allele (P=4.1×10^{-4}). The effect of the *TCF7L2* variant on the BS variance of glucose levels was estimated to be 5.7% increase per allele. Therefore, the correlation between relative pairs was increased by 11.3% per allele, showing that the variant is increasing estimated heritability based on correlation between relative pairs.

5.2 Paper II - A GWAS on structural measures of the corneal endothelium

The second paper of this theses reports results of a GWAS on four quantitative corneal measures; corneal endothelial cell density (cells/mm^2), coefficient of cell size variation (CV), percentage of hexagonal cells (HEX) and central corneal thickness (CCT).

Corneal endothelial images of 6,125 Icelanders were analyzed in the study, obtained as a part of the DHS (see section 4.1.3).

5.2.1 Sex and age effect on corneal traits

Corneal endothelial cells are incapable of mitosis and the number of cells in the endothelium decreases with age. Due to the loss of cells, the remaining cells grow larger and more irregular in shape. Therefore, cell density and HEX decreases with age while CV increases (Figure 11). We also tested if these traits were different between the sexes and found that women have considerably lower HEX values than men (48.3% vs 50.6%, P=8.4×10^{-43}), while women have slightly higher CV and cell density (30.3 vs 29.6, P=2.6×10^{-6}, and 2663 vs 2639 cells/mm^2, P=1.8×10^{-3}). As has been reported before, CCT values do not associate with age, and men have higher values than women on average (565.4 vs 561.5 μm, P=1.9×10^{-4}).

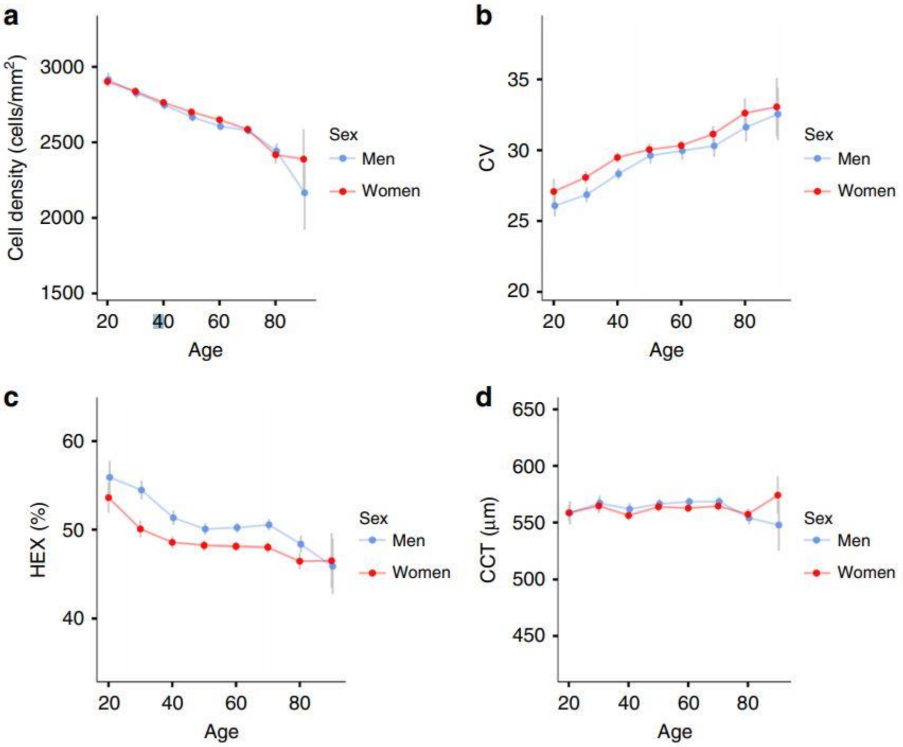

Figure 11. *The mean of the trait by age groups for women in red and men in blue. The gray lines show 95% confidence intervals.*

5.2.2 Study design

We tested 34 million sequence variants (section 4.2.1) for association under the additive model with the four corneal traits. Ten sequence variants were genome-wide significant according to the significance thresholds that are dependent on sequence variant annotation (Figure 12). Seven of the variants are novel and two of them associate most strongly with cell density. Three other variants associated with CV and five variants associated with CCT. The estimated correction factors from the LD score regression were 1.05, 1.03, 1.03 and 1.06 for cell density, CV, HEX, and CCT, respectively.

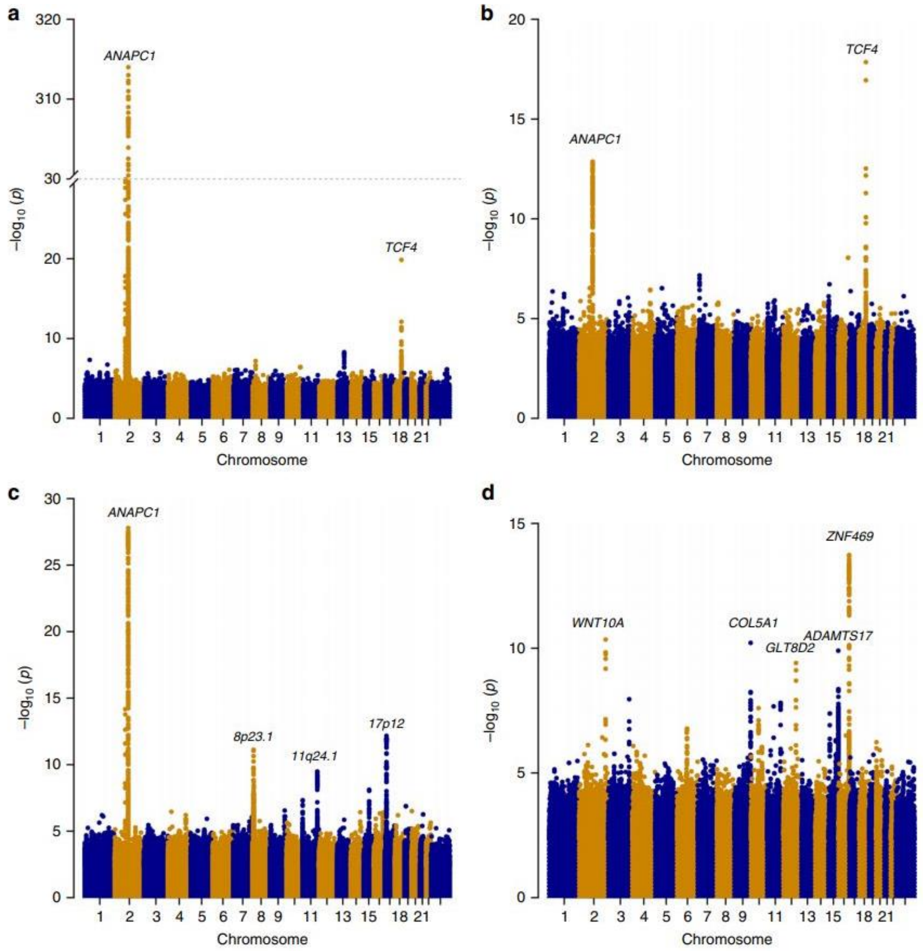

Figure 12. *Manhattan plots showing the association results for the GWAS of the four corneal traits. The $-\log_{10}$ P-values are shown for each variant against their chromosomal position.*

We further explored the associating variants, by assessing their effect on several ocular biomechanics such as corneal hysteresis (CH), corneal resistance factor (CRF), Goldmann correlated intraocular pressure (IOPg), and corneal compensated intraocular pressure (IOPcc). Additionally, we assessed their effect on ocular diseases such as glaucoma and corneal dystrophies.

5.2.3 GWAS results

In this thesis, we will focus on the key results regarding the two variants that associate with cell density (see Paper II for further results).

Two sequence variants associated with cell density; an intronic variant, rs78658973[A], located 0.4 kb downstream of *ANAPC1* and a microsatellite in *TCF4*.

The strongest association is represented by the intronic variant, rs78658973[A], near *ANAPC1* that associates with decreased cell density (β=−0.77 SD, P=1.8×10^{-314}, MAF=28.3%). The variant is highly correlated with 113 variants in the region, but none of them are protein coding. The variant also associates with CV (β=0.23 SD, P=2.8×10^{-28}) and HEX (β=−0.16 SD, P=2.6×10^{-19}). Since CV and HEX correlate with cell density (r=-0.33 and r=0.15), the effects of the variant on CV and HEX are mostly driven by the strong association between the variant and cell density. When we adjust for cell density, the effects on CV and HEX are no longer significant (CV adjusted for cell density: $\beta = -0.04$ SD, P = 0.049; HEX adjusted for cell density: $\beta = -0.04$ SD, P = 0.072).

The effect of rs78658973[A] on cell density is unusually large for such a complex trait, and the fraction of variance of cell density explained by this one variant is 24%. The average cell density per age and genotype groups are shown in Figure 13). Homozygous carriers of rs78658973[A] have on average 455 fewer cells per mm^2 compared to non-carriers and the mean cell density for 30-year old homozygous carriers is lower than the mean cell density for 70-year old non-carriers.

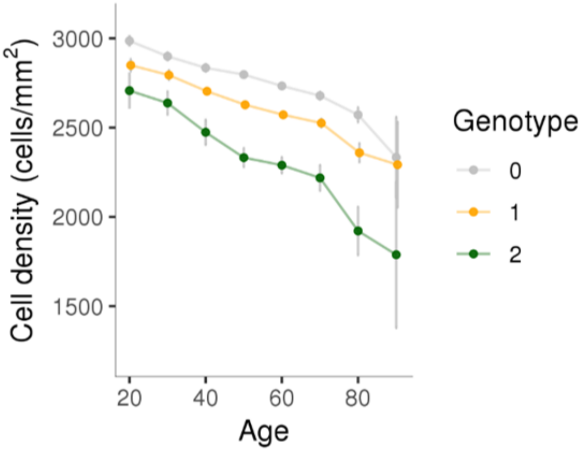

Figure 13. *The mean cell density per age group for non-carriers of the ANAPC1 variant in grey, heterozygous carriers in yellow and homozygous carriers in green.*

The other variant that associates with cell density is a microsatellite in *TCF4*, where the effect allele is defined as a CTG repeat of length at least 33 (MAF=6.1%) and corresponds to the expanded CTG 18.1 allele (OMIM: 602272, allelic variant 0.0007). This microsatellite is a known pathogenic variant, according to Clinvar, reported to cause autosomal dominant FECD[56–58]. FECD is a disease of the corneal endothelium and is the most common indication for corneal transplantation[20,59]. It affects around 4% of people over 40 years old (OMIM: 602272) and is characterized by premature loss of endothelial cells resulting in increased variability in cell shape and size which can cause corneal edema and visual loss. Consistent with these characteristics of the disease, the expanded CTG 18.1 allele associates with lower cell density and HEX ($\beta = -0.38$ SD, P = 1.6 × 10^{-19} and $\beta = -0.37$ SD, P = 5.9 × 10^{-18}, respectively).

5.2.4 The effect of cell density associating variants on diseases and other corneal metrics

To estimate the effect of the cell density associating variants on ocular diseases, we performed a meta-analysis of the Icelandic and UK Biobank data using ICD codes for glaucoma and corneal diseases. The results for the variants at *ANAPC1* and *TCF4* are shown in Table 5. We replicate the known effect of the microsatellite in *TCF4* on FECD using the ICD code for hereditary corneal dystrophies (OR=7.8, P=3.3×10^{-31}). However, the variant near *ANAPC1* does not associate with any of the diseases.

Table 5. *The association of the TCF4 and ANAPC1 variants with ocular diseases in a meta-analysis of GWAS results from Iceland and the UK Biobank. The number of cases and controls are shown for each disease.*

		N		*TCF4*		*ANAPC1*	
Trait	**ICD10 codes**	**Cases**	**Controls**	**P-value**	**OR**	**P-value**	**OR**
Disorders of cornea	H18	1,236	663K	0.014	1.78	0.43	0.95
Corneal degeneration	H18.4	199	684K	**9.9×10^{-9}**	**3.98**	0.86	0.98
Hereditary corneal dystrophies	H18.5	330	684K	**3.3×10^{-31}**	**7.77**	0.030	0.81
Glaucoma	H40	8,432	641K	0.15	0.92	0.048	1.04
Primary open angle glaucoma (POAG)	H40.1	2,296	706K	0.51	0.94	0.24	0.96
Primary angle closure glaucoma (PACG)	H40.2	777	637K	0.24	0.61	0.19	0.92

The participants of the DHS also had various ocular biomechanics measured, including corneal hysteresis (CH), corneal resistance factor (CRF), Goldmann correlated intraocular pressure (IOPg) and corneal compensated intraocular pressure (IOPcc).

Interestingly, the *ANAPC1* and *TCF4* variants both associate strongly with CH (β=0.19 SD, P=2.6×10^{-19} and β=−0.29 SD, P=3.1×10^{-12}). The alleles that associate with decreased cell density have different direction of effect on CH, i.e. the *ANAPC1* variant associates with increased CH while the *TCF4* variant associates with decreased CH. CH is a measure of the cornea's ability to absorb and dissipate energy[60,61] and low CH has been associated with faster rate of glaucoma progression[62–64]. However, the variants do not associate with glaucoma in our data (Table 5).

5.2.5 Glaucoma and corneal measures

Of the DHS participants, 3.2% have glaucoma according to self-reported information. We estimated the effect of glaucoma disease status on the corneal measures and observed that CH and cell density are most strongly associated with glaucoma (Table 6) such that glaucoma patients have fewer cell in the endothelium and lower CH.

Table 6. *The first two columns show the effect of glaucoma status on corneal traits, adjusting for age and sex. The other two columns show the correlation (R^2) between the effect of glaucoma variants on glaucoma and their effect on the corneal traits.*

	Glaucoma status		**Glaucoma variants**	
	β (SD)	**P-value**	**R^2**	**P-value**
CH	-0.37	$3.8 \cdot 10^{-7}$	0.03	0.51
CD	-0.35	$2.1 \cdot 10^{-6}$	0.03	0.49
IOPcc	0.29	$7.3 \cdot 10^{-5}$	0.31	0.03
CCT	-0.25	$4.7 \cdot 10^{-4}$	0.02	0.58
IOPg	0.18	$1.4 \cdot 10^{-2}$	0.22	0.06
HEX	-0.16	$3.0 \cdot 10^{-2}$	0.02	0.57
CV	0.13	$6.9 \cdot 10^{-2}$	0.09	0.27
CRF	-0.07	0.34	0.02	0.63

Because of these associations between glaucoma status and corneal measures, we estimate the effect of previously published glaucoma variants that replicate in our data on corneal measures (Figure 14). Despite the correlation between glaucoma and the corneal measures, the correlation between the glaucoma variants effect on glaucoma and their effect on the corneal traits was not significant for any trait (Table 6). This suggests that these glaucoma variants do not confer their risk of glaucoma through their effect on CH and cell density.

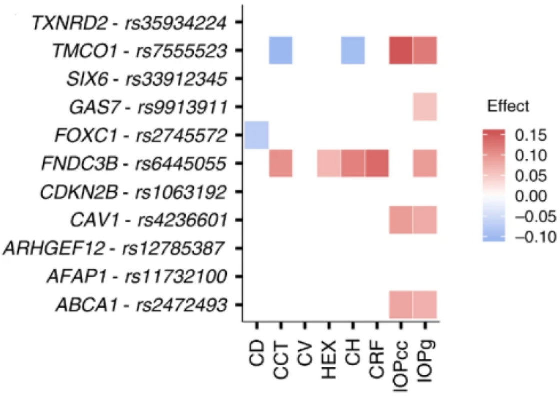

Figure 14. *Effect of previously reported glaucoma variants on corneal traits for the glaucoma risk increasing allele. Effects on the traits are shown for significant associations after adjusting for multiple testing with the Benjamini–Hochberg false discovery rate procedure for each variant. Effect are shown in red for positive effects and blue for negative effects.*

5.3 Paper III – A GWAS meta-analysis of age-related hearing impairment

The third paper of this thesis reports results from a meta-analysis of three GWAS on ARHI, using audiometric measures from two non-overlapping Icelandic datasets (DHS and NIHSI) and self-reported hearing difficulty from the UK Biobank (see section 4.1.4).

5.3.1 Demographics of ARHI in Iceland

The DHS and NIHSI datasets are both based on audiometric measures. The DHS dataset was obtained as a part of a broad phenotype collection for a general population sample, while at the NIHSI, most of the subjects that have audiometric measurements have problems with their hearing (Figure 15).

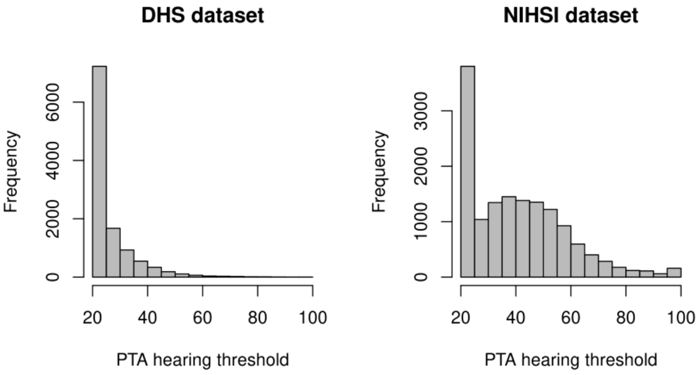

Figure 15. *Histograms of PTA hearing thresholds from* ***a)*** *DHS and* ***b)*** *NIHSI datasets.*

The DHS dataset therefore provides a good opportunity to analyze the prevalence of ARHI in Iceland, which was 36.1% for mild (PTA>25), 7.7% for moderate (PTA>40), 1.1% for severe (PTA>60) and 0.1% for profound (PTA>80) impairment. The prevalence of ARHI increases rapidly with age and ARHI is more common for men than women (Figure 16).

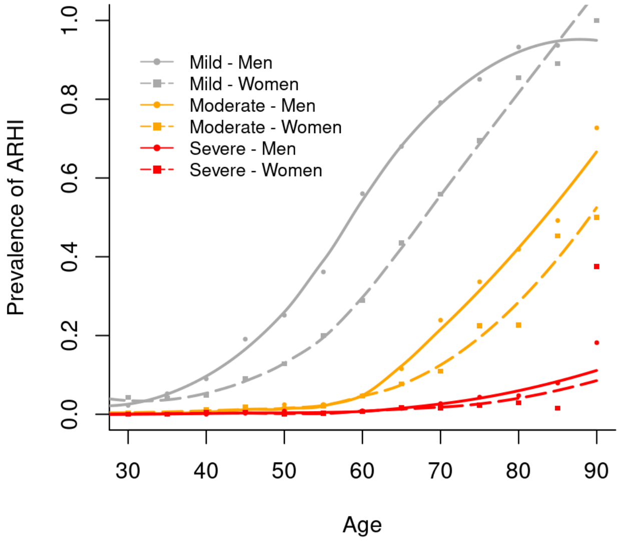

Figure 16. *The prevalence of hearing impairment by age groups in the DHS dataset. Shown are the fraction of women (squares) and men (round dots) per age group with mild in grey, moderate in yellow and severe hearing impairment in red.*

5.3.2 Study Design

In total, we tested 47 million variants for association with ARHI in a meta-analysis of three GWAS (Figure 17) using both the additive and recessive models. The estimated correction factors for ARHI were 1.05, 1.20 and 1.05, under the additive model, and 1.01, 1.09 and 1.00, under the recessive model, in DHS, NIHSI and UKB datasets respectively.

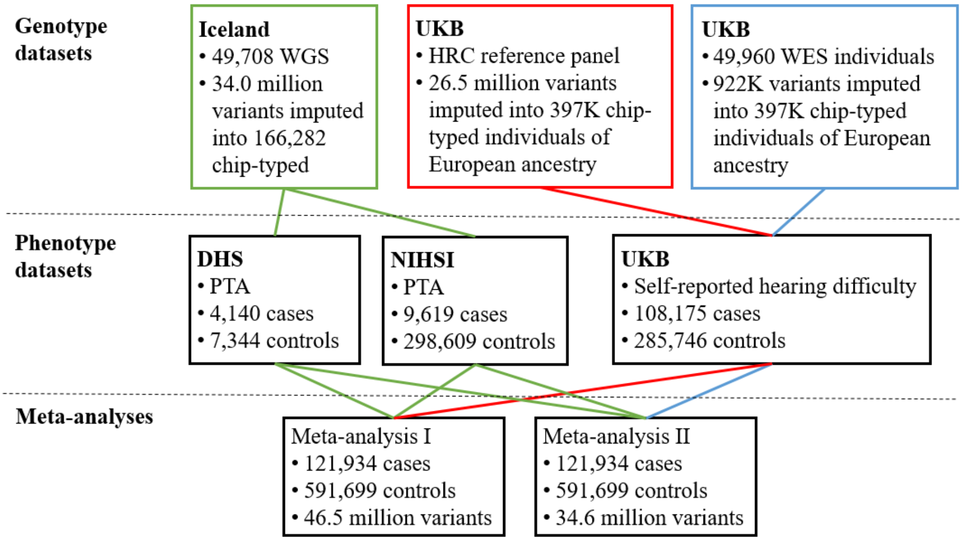

Figure 17. *A flowchart explaining the study design of the ARHI GWAS meta-analysis. The UK Biobank (UKB) GWAS was performed on two genotype datasets labelled in red and blue. WGS = Whole genome sequenced, HRC = Haplotype Reference Consortium, WES = Whole exome sequenced, PTA = Pure tone average, DHS = deCODE health study, NIHSI = National Institute of Hearing and Speech in Iceland.*

Fifty-five independent variants at 48 loci satisfied our genome-wide significance thresholds that are dependent on sequence variant annotation (Figure 18).

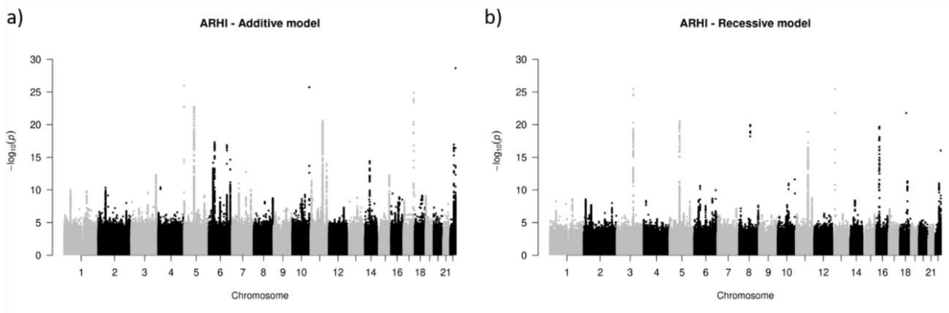

Figure 18. *Manhattan plots showing the association results for the ARHI meta-analysis under the **a)** the additive model and **b)** the recessive model. The $-log_{10}$ P-values are shown for each variant against their chromosomal position.*

If a variant has MAF=f, the expected genotype frequency (EGF) is 2f(1-f) for heterozygous carriers and f^2 for homozygous carriers. In this paper, we define rare variants as variants with EGF below 1%, i.e. variants detected with the additive model with MAF<0.5% and variants detected with the recessive model with MAF<10%. Twenty-one of the detected variants are rare. Due to the PTA based definition of ARHI cases in the Icelandic datasets

(PTA>25 dB HL), individuals that are completely deaf, or have prelingual or childhood-onset hearing loss, are not excluded from the analysis. Therefore, the GWAS might detect variants that are causing deafness instead of ARHI. Because of this, for each rare variant we detected, we estimated whether they are associating with ARHI, by fitting a linear model using the PTA hearing threshold for the carriers as a response and age as a covariate. Four variants had predicted PTA over 25 dB HL at 10 years of age, and we therefore considered them to be causing childhood-onset hearing loss. Thus, 51 variants associated with AHRI, 41 under the additive model and 10 under the recessive model. We furthermore used a gene-based burden test, where LOF variants with MAF<2% were aggregated and tested together, yielding one additional ARHI gene; *AP1M2*.

We explored the associating ARHI variants, by assessing their effect on tinnitus and on ARHI for each frequency; 0.5, 1, 2, 4, 6 and 8 kHz. We additionally analyzed their effects on ARHI for heterozygous and homozygous carriers separately by using the full genotypic model. Furthermore, we constructed a GRS for ARHI using the effect estimates from the UK Biobank and assessed its predictive abilities in the Icelandic datasets.

5.3.3 GWAS results

In this thesis, we will briefly report the key results (see Paper III for further results).

Twenty-one of the ARHI variants are novel associations, and 16 are rare variants with EGF less than 1%. Five of the rare variants are in genes that have not been reported for hearing before, six are in Mendelian deafness genes and two are secondary signals at ARHI loci (Figure 19).

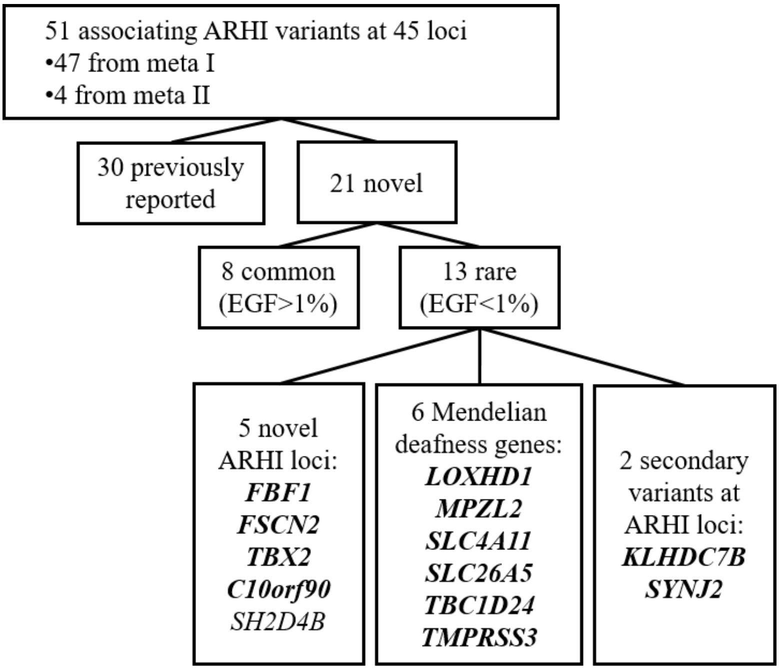
51 associating ARHI variants at 45 loci
•47 from meta I
•4 from meta II
30 previously reported
21 novel
8 common (EGF>1%)
13 rare (EGF<1%)
5 novel ARHI loci: FBF1 FSCN2 TBX2 C10orf90 SH2D4B
6 Mendelian deafness genes: LOXHD1 MPZL2 SLC4A11 SLC26A5 TBC1D24 TMPRSS3
2 secondary variants at ARHI loci: KLHDC7B SYNJ2

Figure 19. *A flowchart showing a summary of the results for the ARHI GWAS meta-analysis.*

In Iceland, 4.9% of the population are carriers of at least one of the 16 rare ARHI variants. Carriers of rare variants have a 2.2-fold ($P=1.0\times10^{-12}$) greater risk of mild hearing impairment compared to the rest of the population, 3.0-fold ($P=1.4\times10^{-9}$) greater risk of moderate impairment and 5.6-fold ($P=1.9\times10^{-8}$) greater risk of severe impairment (Figure 20).

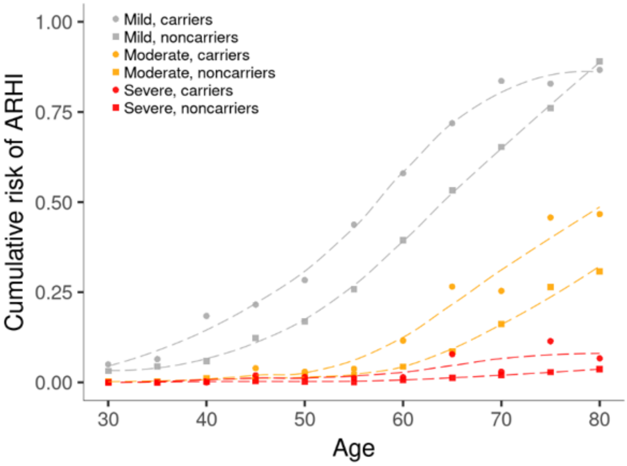

Figure 20. *The cumulative risk of mild, moderate and severe hearing impairment in the DHS dataset among the 4.9% of subjects that are carriers of any of the 16 rare ARHI variants (round dots) and the 95% that are not carriers (squares).*

Of the 16 rare variants, 15 are in protein coding regions and 13 are novel. Of the novel variants, a missense variants in *LOXHD1* and a tandem duplication in *FBF1* associate most strongly with ARHI (OR=3.7, P=1.7×10^{-22} and OR=4.2, P=5.7×10^{-27}, respectively).

The missense variant, p.Arg1090Gln, in *LOXHD1* (MAF$_{\text{Iceland}}$=2.96% and MAF$_{\text{UK}}$=1.99%) was detected using the recessive model and is one of the six rare ARHI variants that are located in a Mendelian deafness genes. Out of the 62 homozygous carriers with audiometric measurements in the Icelandic datasets, 82.3% have at least mild hearing impairment and 48.4% have moderate to profound hearing impairment (Figure 21.a). The rare tandem duplication in *FBF1* (MAF$_{\text{Ice}}$=0.22%) was detected in Iceland under the additive model and spans 7,282 base pairs and covers exons 4 to 7. For the duplication, 81.5% of 162 carriers with audiometric measurements have at least mild hearing impairment and 48.4% have moderate to profound hearing impairment (Figure 21.b).

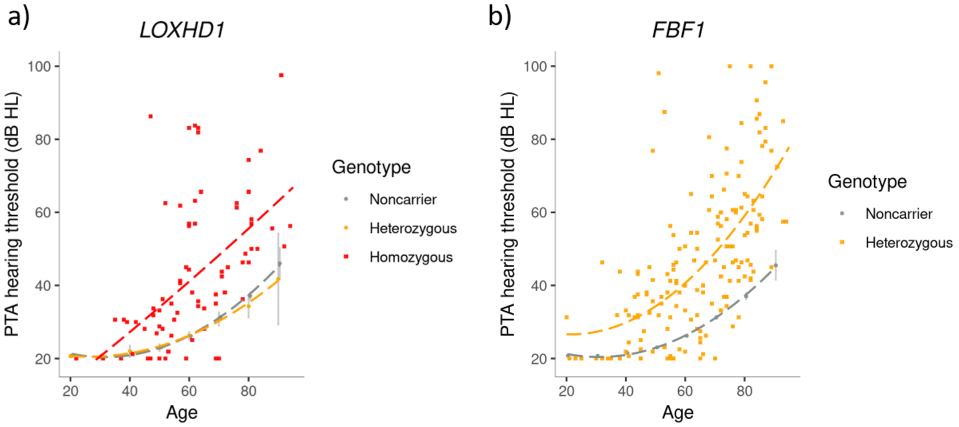

Figure 21. *Changes in PTA hearing thresholds by age for carriers of the variants in* ***a)*** *LOXHD1 and* ***b)*** *FBF1. In figure a, PTA means of non-carriers in the DHS dataset are represented with grey dots, means of heterozygotes in DHS dataset with orange dots and the PTA hearing threshold of the homozygous carriers in DHS and NIHSI datasets are indicated by red squares. In figures b PTA means of non-carriers in the DHS dataset are represented with grey dots and the PTA hearing threshold of the heterozygous carriers in DHS and NIHSI datasets are indicated by yellow squares.*

5.3.4 Genetic risk scores predict ARHI risk

We constructed a GRS using the 35 common variants (EGF>1%) and effect estimates from the UK Biobank as weights and taking into account if the variants associated under the additive or recessive models (section 4.6). The GRS associates with ARHI in both Icelandic datasets (OR=1.31, P=4.1×10^{-29} and OR=1.18, P=7.5×10^{-39} in DHS and NIHSI datasets respectively) and the risk of ARHI for individuals increases over GRS deciles (Figure 22.a). Individuals in the top GRS decile have a 2.5-fold (P=6.1×10^{-18}) greater risk of ARHI relative to those in the bottom decile in the DHS dataset. Furthermore, when looking at the fraction of ARHI cases for 5-year age groups, we see that individuals in the top GRS decile develop ARHI on average 10 years younger than those in the bottom decile (Figure 22.b).

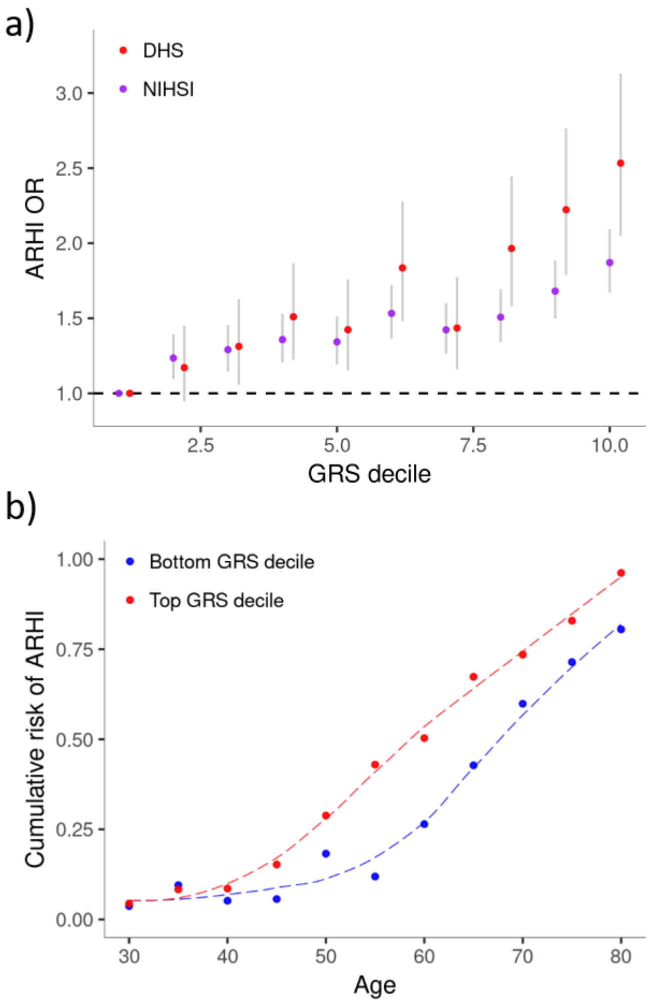

Figure 22. a) *The ORs for ARHI for each GRS decile in the DHS and NIHSI datasets relative to the bottom decile.* ***b)*** *The cumulative risk of ARHI among subjects in the DHS dataset in the bottom GRS decile in blue and the top GRS decile in red.*

If we compare the 4.9% that carry at least one of the 16 rare ARHI variant to the bottom GRS decile, the ORs are 3.4 for mild, 6.1 for moderate and 9.2 for severe hearing impairment (P=3.0×10^{-19}, 8.4×10^{-13} and 1.0×10^{-7}, respectively). Thus, relative to the bottom GRS decile, the risk of ARHI for carriers of rare ARHI variants is comparable to the risk of ARHI of individuals in top GRS decile (P heterogeneity=0.075), while the risk of moderate and severe hearing impairment for carriers of rare variants is substantially greater than the risk for the top GRS decile (P heterogeneity<0.05).

5.3.5 Association of ARHI variants with ARHI under the full genotypic model

Ten out of the 51 ARHI variants, were detected under the recessive model. To further explore the effect of heterozygous and homozygous carriers separately on ARHI, we fit a full genotypic model for each of the ARHI variants using two parameters for heterozygotes and homozygotes. Four variants in *TYR*, *KLHDC7B*, *SYNJ2* and *CLRN2*, that were detected using the additive model, had stronger effect of homozygous carriers than expected by the additive model ($P<0.05$, Table 7). Three variants in *ILDR1*, *CHMP4C* and *CCDC68* that we detected using the recessive model, were reported by Wells et al. using the additive model. The results using the genotypic model shows that these variants are truly associating with ARHI under recessive mode of inheritance, showing no significant effects for heterozygous carriers (Table 7).

Table 7. *The effect of seven ARHI variants on ARHI per genotype, four that had stronger effect on homozygous carriers than expected by the additive model, and three recessive variants that have been reported under the additive model. The deviation column shows the P-value when comparing the full genotypic model to the additive model.*

Chr	Position	Gene	MAF	Full genotypic model P-value	Heterozygote		Homozygote		P deviation
					Effect	P-value	Effect	P-value	
3	121993204	*ILDR1*	29.72	$9.2 \cdot 10^{-19}$	1.01	0.40	1.13	$8.5 \cdot 10^{-20}$	$4.6 \cdot 10^{-15}$
8	81753241	*CHMP4C*	6.83	$3.6 \cdot 10^{-16}$	1.03	$2.3 \cdot 10^{-2}$	1.53	$7.9 \cdot 10^{-15}$	$2.1 \cdot 10^{-14}$
22	50549676	*KLHDC7B*	3.57	$1.2 \cdot 10^{-33}$	1.13	$1.8 \cdot 10^{-20}$	1.93	$1.0 \cdot 10^{-11}$	$1.9 \cdot 10^{-6}$
6	158071628	*SYNJ2*	0.37	$3.3 \cdot 10^{-14}$	1.29	$8.3 \cdot 10^{-11}$	22.61	$4.3 \cdot 10^{-5}$	$2.2 \cdot 10^{-4}$
18	54937957	*CCDC68*	21.12	$3.1 \cdot 10^{-9}$	1.01	$6.8 \cdot 10^{-2}$	1.11	$9.4 \cdot 10^{-8}$	$5.4 \cdot 10^{-4}$
11	89284793	*TYR*	30.15	$4.5 \cdot 10^{-20}$	1.03	$2.3 \cdot 10^{-5}$	1.13	$7.6 \cdot 10^{-12}$	$2.3 \cdot 10^{-3}$
4	17522947	*CLRN2*	12.79	$4.2 \cdot 10^{-11}$	1.04	$3.5 \cdot 10^{-6}$	1.16	$6.6 \cdot 10^{-5}$	$2.8 \cdot 10^{-3}$

5.3.6 ARHI and tinnitus

We estimated the effect of the ARHI variants on tinnitus by using self-reported information from DHS and UK Biobank (N_{cases}=47,657 and $N_{controls}$=111,607). For ARHI variants that were detected using the additive model, we estimated their effect on tinnitus with the additive model. Similarly, for variants that were detected under the recessive model, we used the recessive model when estimating their effect on tinnitus. Thirteen ARHI variants also associated with tinnitus when controlling the false-discovery rate at 0.05 using the Benjamini-Hochberg procedure, and for all of them, the ARHI risk increasing allele associated with increased risk of tinnitus. For all ARHI variants, their effect on ARHI was highly correlated with their effect on tinnitus (Figure 23).

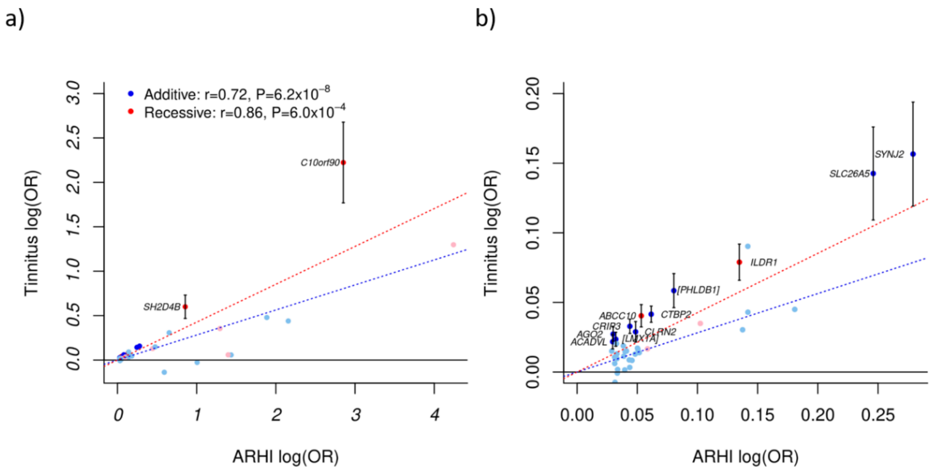

Figure 23. *The effect of the ARHI variants on ARHI is plotted against their effect on tinnitus for* ***a)*** *all ARHI variants and* ***b)*** *zoomed-in on variants with ARHI OR <1.35. ARHI variants detected under the additive model are colored blue and ARHI variants detected under the recessive model are colored red. Variants that affect tinnitus, controlling the false discovery rate at 0.05, are plotted with darker color and labelled with their corresponding gene. The effects are shown for the ARHI risk increasing allele. Error bars represent 95% confidence intervals. The dotted lines represent results from a weighted linear regression using MAF(1-MAF) as weights, red for recessive variants and blue for additive, and the weighted correlation coefficients (r) and the corresponding P-values are shown in figure a.*

6 Discussion

6.1 Conclusions

6.1.1 Paper I - Sequence variants affecting the variance of glucose levels

In this study, we implemented variance association models to estimate the effect of sequence variants on the BS and WS variance of quantitative traits. We found that variants in *TCF7L2*, *GCK*, *G6PC2* and *GRB10*, that are known to associate with mean glucose levels, also associate with the variance of glucose levels. The observation is robust to removing diabetics and individuals on diabetic medication from the dataset. We observed that variants that affect both the mean and the BS variance of glucose levels increase T2D risk more than variants that increase mean levels but reduce the BS variance. Furthermore, we found that the effect of variants on the BS variance of glucose levels are as important for genetic risk prediction of T2D as the effect of variants on the mean. Apart from increasing our understanding of the impact of genetics on glucose metabolism and control, this observation helps resolve the question of why sequence variants that associate with higher fasting glucose levels do not always associate with increased risk of T2D.

The effects of these variants on the variance of glucose are so large that they are certain to be an important component in the creation of heritability. Indeed, we show that the covariance between first degree relative pairs is proportional to their variance effect and that for the *TCF7L2* variant, the correlation between close relatives is greater for carriers than non-carriers. Thus, the effect of these variants on the variance contribute substantially to the missing heritability of glucose levels.

6.1.2 Paper II - A GWAS on structural measures of the corneal endothelium

In this study we described the first GWAS on structural measurements of the corneal endothelium including cell density, coefficient of cell size variation and percentage of hexagonal cells for which we observed several associating variants.

Healthy corneal endothelium is necessary for visual perception and its dysfunction is the most common reason for corneal transplantation. A minimum number of endothelial cells is needed to maintain proper hydration of the cornea and lower cell density is found in patients with various eye diseases such as glaucoma and corneal dystrophies. Interestingly, we discovered a strong association between lower cell density and a sequence variant near *ANAPC1* that accounts for a quarter of the population variance of cell density which is extremely high for such a complex human trait. No other variant in the GWAS catalog is close to explaining such a high fraction of variance of its associating quantitative phenotype, when the trait is not a direct measurement of the protein affected by the variant.

Despite correlations between cell density and ocular diseases, such as corneal dystrophies and glaucoma, the *ANAPC1* variant does not associate with risk of eye diseases in our data. That shows that endothelial cell density is largely under genetic control without affecting the risk of disease. This finding is clinically relevant because abnormal loss of endothelial cells is usually the first sign of endothelial diseases and prior to intraocular surgery, cell density is used for assessing the risk of endothelial failure.

6.1.3 Paper III – A GWAS meta-analysis of ARHI

In this study we described the largest genome-wide association meta-analysis to date on age-related hearing impairment (ARHI), using audiometric measurements from two non-overlapping Icelandic datasets and information on self-reported hearing difficulty from the UK Biobank. We observed 21 novel sequence variants associating with ARHI, using both the additive and recessive models. Previous GWAS on ARHI have reported on common variants with small to moderate effects, while in this study, 13 of the novel variants have rare genotypes with large effects. Five of the novel rare variants are at loci that have not been reported for hearing before, one of which is a tandem duplication that covers exons 4 to 7 of *FBF1*.

We constructed an ARHI GRS from common variants using effect sizes from the UK Biobank. We showed that using the GRS to stratify individuals into risk groups, can identify individuals at risk comparable to carriers of rare ARHI variants with high penetrance. However, the rare ARHI variants seem to cause more severe ARHI than the common variants.

We explored the effect of the ARHI variant on tinnitus, a phenotype known to be correlated with ARHI. We found a high correlation between the effect sizes of ARHI variants on ARHI and tinnitus suggesting that these traits share genetic causes.

Interestingly, 20% of the detected variants associated more strongly under the recessive model, which is unusual for age-related diseases. The missense variant in *LOXHD1*, which is genome-wide significant in the UK Biobank dataset, was not detected under the additive model in a previous study using the same dataset. Furthermore, variants that have been previously reported with additive effects, are truly associating with ARHI under recessive mode of inheritance. This illustrates the importance of also using the recessive model when searching for variants affecting the risk of ARHI.

6.2 Future perspectives

6.2.1 Paper I - Sequence variants affecting the variance of glucose levels

In Paper I we implemented a variance model to detect the effect of variants on the BS and WS variance of quantitative traits. The variance models have some drawbacks and could be developed further in future research. For instance, the likelihood has a closed form solution if the genotype is 0, 1 or 2, but does not take into account genotype probabilities. Another issue is that an undetected secondary variant can lead to spurious variance association at the primary variant. In Paper I, we examined secondary variant at the loci that associated with the variance of glucose levels and found that they had no impact on the variance effects. But in future studies, the model could be developed so that it accounts for the mean effects of secondary variants.

In Paper I, we used the variance models to estimate the variance effects of known glucose variants on glucose levels. We also performed a genome-wide scan and found no additional variants associating with the variance of glucose levels. In this thesis we constricted the analysis to glucose levels, but there is great potential in exploring variance effects on other quantitative traits. Since larger sample sizes are needed to detect variance effects than mean effects, the advent of biobanks that have genetic data combined with high-dimensional phenotypic data obtained for a large number of subjects, provide an opportunity to analyze variance effects for several quantitative traits that can enhance our understanding of the biology and genetics of those traits as well as account for some of the missing heritability.

6.2.2 Papers II and III – GWAS on corneal traits and age-related hearing impairment

In papers II and III, GWAS were performed to search for variants associating with sensory traits. Future research has the potential to further increase the understanding of the genetics of these traits.

In Paper II, corneal measures obtained from a specular microscopy were analyzed that have not been analyzed in GWAS before. Increasing sample size would most likely lead to more associating variants that would further increase the knowledge about these corneal traits. Since Paper II was written, the sample size of the deCODE health study has almost doubled, and we already have an example of a missense variant in *SLC4A11* that we found to associate with ARHI in Paper III, that also associates genome-wide significantly with endothelial cell density but was not detected at the time when Paper II was written.

In Paper III, the largest to date GWAS meta-analysis on ARHI was performed. Increasing sample sizes by continuing collaboration of different study groups could further increase the predictive abilities of the GRS and likely identify even more rare variants associating with ARHI.

Furthermore, GWAS studies on these traits and in general, will benefit from improved quality in genotype data by continuing to whole-genome sequence individuals on a large-scale basis as well as better genotyping methods. For instance, a novel method to genotype structural variants[65] was used in the Icelandic datasets in Paper III, which led to the detection of the strongest associating ARHI variant in Iceland, a tandem duplication in *FBF1*. Applying this method on other datasets, such as the UK Biobank dataset, might result in more novel findings.

References

1. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47,** 435–444 (2015).

2. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4,** 170115 (2017).

3. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48,** 314–317 (2016).

4. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* (2009). doi:10.1038/nature08625

5. Wei, W.-H. *et al.* Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. *Sci. Rep.* **6,** 25014 (2016).

6. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490,** 267–272 (2013).

7. Topless, R. K. *et al.* Association of SLC2A9 genotype with phenotypic variability of serum urate in pre-menopausal women. *Front. Genet.* **6,** 1–9 (2015).

8. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. *PLoS Genet.* **6,** 1–10 (2010).

9. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0225-6

10. Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* (2019). doi:10.1126/sciadv.aaw3538

11. Levene, H. Robust tests for equality of variances. *Contrib. to Probab. Stat. Essays* (1960).

12. Shen, X., Pettersson, M., Rönnegård, L. & Carlborg, Ö. Inheritance Beyond Plain Heritability: Variance-Controlling Genes in Arabidopsis thaliana. *PLoS Genet.* **8,** (2012).

13. Rönnegård, L. & Valdar, W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics* (2011). doi:10.1534/genetics.111.127068

14. Cao, Y., Wei, P., Bailey, M., Kauwe, J. S. K. & Maxwell, T. J. A versatile omnibus test for detecting mean and variance heterogeneity. *Genet. Epidemiol.* (2014). doi:10.1002/gepi.21778

15. Falconer, D. S. & Mackay, T. F. C. Introduction to quantitative genetics. *Introduction to quantitative genetics* **4,** 43 (1996).

16. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* (2003). doi:10.1093/ije/dyg070

17. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **46,** 234–44 (2012).

18. Ruth, N. & Peralta, V. 2015 Eye Banking Statistical Report. *Eye Bank Assoc. Am. Washington, DC* **3,** 58–69 (2016).

19. Bourne, W. M. Biology of the corneal endothelium in health and disease. *Eye* **17,** 912–918 (2003).

20. Sarnicola, C., Farooq, A. V. & Colby, K. Fuchs Endothelial Corneal Dystrophy: Update on Pathogenesis and Future Directions. *Eye Contact Lens* **0,** 1–10 (2018).

21. Joyce, N. Proliferative capacity of corneal endothelial cells. *Exp. Eye Res.* **95,** 16–23 (2012).

22. Gao, X. *et al.* Genome-wide association study identifies WNT7B as a novel locus for central corneal thickness in Latinos. *Hum. Mol. Genet.* **25,** 5035–5045 (2016).

23. Li, X. *et al.* Genetic association of COL5A1 variants in keratoconus patients suggests a complex connection between corneal thinning and keratoconus. *Investig. Ophthalmol. Vis. Sci.* **54,** 2696–2704 (2013).

24. Cuellar-Partida, G. *et al.* WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Hum. Mol. Genet.* **24,** 5060–5068 (2015).

25. Igo, R. P. *et al.* Differing Roles for TCF4 and COL8A2 in Central Corneal Thickness and Fuchs Endothelial Corneal Dystrophy. *PLoS One* **7,** e46742 (2012).

26. Yi, L. *et al.* Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* **45,** 155–163 (2013).

27. Iglesias, A. I. *et al.* Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nat. Commun.* **9,** 1864 (2018).

28. Vitart, V. *et al.* New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. *Hum. Mol. Genet.* **19,** 4304–4311 (2010).

29. Afshari, N. A. *et al.* Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat. Commun.* **8,** 14898 (2017).

30. Gagnon, M. M., Boisjoly, H. M., Brunette, I., Charest, M. & Amyot, M. Corneal endothelial cell density in glaucoma. *Cornea* **16,** 314–8 (1997).

31. Morton, C. C. & Nance, W. E. Newborn hearing screening - A silent revolution. *New England Journal of Medicine* (2006). doi:10.1056/NEJMra050700

32. Frolenkov, G. I., Belyantseva, I. A., Friedman, T. B. & Griffith, A. J. Genetic insights into the morphogenesis of inner ear hair cells. *Nature Reviews Genetics* **5,** 489–498 (2004).

33. Friedman, R. A. *et al.* GRM7 variants confer susceptibility to age-related hearing impairment. *Hum. Mol. Genet.* **18,** 785–796 (2009).

34. Fransen, E. *et al.* Genome-wide association analysis demonstrates the highly polygenic character of age-related hearing impairment. *Eur. J. Hum. Genet.* **23,** 110–115 (2015).

35. Girotto, G. *et al.* Hearing function and thresholds: A genome-wide association study in European isolated populations identifies new loci and pathways. *J. Med. Genet.* **48,** 369–374 (2011).

36. Van Laer, L. *et al.* A genome-wide association study for age-related hearing impairment in the Saami. *Eur. J. Hum. Genet.* **18,** 685–693 (2010).

37. Hoffmann, T. J. *et al.* A Large Genome-Wide Association Study of Age-Related Hearing Impairment Using Electronic Health Records. *PLoS Genet.* **12,** (2016).

38. Vuckovic, D. *et al.* Genome-wide association analysis on normal hearing function identifies PCDH20 and SLC28A3 as candidates for hearing function and loss. *Hum. Mol. Genet.* **24,** 5655–5664 (2015).

39. Wells, H. R. R. *et al.* GWAS Identifies 44 Independent Associated Genomic Loci for Self-Reported Adult Hearing Difficulty in UK Biobank. *Am. J. Hum. Genet.* **105,** 788–802 (2019).

40. Atik, A. Pathophysiology and Treatment of Tinnitus: An Elusive Disease. *Indian Journal of Otolaryngology and Head and Neck Surgery* **66,** 1–5 (2014).

41. Axelsson, A. & Ringdahl, A. Tinnitus-A study of its prevalence and characteristics. *Br. J. Audiol.* **23,** 53–62 (1989).

42. Maas, I. L. *et al.* Genetic susceptibility to bilateral tinnitus in a Swedish twin cohort. *Genet. Med.* **19,** 1007–1012 (2017).

43. Vona, B., Nanda, I., Shehata-Dieler, W. & Haaf, T. Genetics of tinnitus: Still in its infancy. *Frontiers in Neuroscience* **11,** (2017).

44. Azizi, F. *et al.* Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials* **10,** 5 (2009).

45. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12,** e1001779 (2015).

46. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* (2015). doi:10.1038/ng.3247

47. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4,** 170115 (2017).

48. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49,** 1654–1660 (2017).

49. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* https://doi.org/10.1101/166298 (2017). doi:10.1101/166298

50. Hout, C. V. Van *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* 572347 (2019). doi:10.1101/572347

51. Hastie, T. & Tibshirani, R. Generalized Additive Models. *Statistical Science* **1,** 297–310 (1986).

52. Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *J. Natl. Cancer Inst.* **22,** 719–48 (1959).

53. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

54. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46,** 294–298 (2014).

55. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78,** 691–692 (1991).

56. Mootha, V. V., Gong, X., Ku, H. C. & Xing, C. Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in fuchs' endothelial corneal dystrophy. *Investig. Ophthalmol. Vis. Sci.* **55,** 33–42 (2014).

57. Wieben, E. D. *et al.* A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. *PLoS One* **7,** e49083 (2012).

58. Riazuddin, S. A. *et al.* Replication of the TCF4 Intronic variant in Late-onset fuchs corneal dystrophy and evidence of independence from the FCD2 locus. *Investig. Ophthalmol. Vis. Sci.* **52,** 2825–2829 (2011).

59. Adamis, A. P., Filatov, V., Tripathi, B. J. & Tripathi, R. A. mes. C. Fuchs' endothelial dystrophy of the cornea. *Survey of Ophthalmology* **38,** 149–168 (1993).

60. Shah, S., Laiquzzaman, M., Cunliffe, I. & Mantry, S. The use of the Reichert ocular response analyser to establish the relationship between ocular hysteresis, corneal resistance factor and central corneal thickness in normal eyes. *Contact Lens Anterior Eye* **29,** 257–262 (2006).

61. Kotecha, A., Elsheikh, A., Roberts, C. R., Zhu, H. & Garway-Heath, D. F. Corneal thickness- and age-related biomechanical properties of the cornea measured with the ocular response analyzer. *Investig. Ophthalmol. Vis. Sci.* **47,** 5337–5347 (2006).

62. Zhang, C. *et al.* Corneal Hysteresis and Progressive Retinal Nerve Fiber Layer Loss in Glaucoma. *Am. J. Ophthalmol.* **166,** 29–36 (2016).

63. De Moraes, C. V. G., Hill, V., Tello, C., Liebmann, J. M. & Ritch, R. Lower corneal hysteresis is associated with more rapid glaucomatous visual field progression. *J. Glaucoma* **21,** 209–213 (2012).

64. Medeiros, F. A. *et al.* Corneal hysteresis as a risk factor for glaucoma progression: A prospective longitudinal study. *Ophthalmology* **120,** 1533–1540 (2013).

65. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10,** 1–8 (2019).

Part II - Papers

Paper I

Effect of sequence variants on variance in glucose levels predicts type 2 diabetes risk and accounts for heritability

Erna V Ivarsdottir[1,2], Valgerdur Steinthorsdottir[1], Maryam S Daneshpour[3], Gudmar Thorleifsson[1], Patrick Sulem[1], Hilma Holm[1], Snaevar Sigurdsson[1], Astradur B Hreidarsson[4], Gunnar Sigurdsson[4], Ragnar Bjarnason[5,6], Arni V Thorsson[5], Rafn Benediktsson[4,6], Gudmundur Eyjolfsson[7], Olof Sigurdardottir[8], Isleifur Olafsson[9], Sirous Zeinali[10], Fereidoun Azizi[11], Unnur Thorsteinsdottir[1,6], Daniel F Gudbjartsson[1,2] & Kari Stefansson[1,6]

Sequence variants that affect mean fasting glucose levels do not necessarily affect risk for type 2 diabetes (T2D). We assessed the effects of 36 reported glucose-associated sequence variants[1] on between- and within-subject variance in fasting glucose levels in 69,142 Icelanders. The variant in *TCF7L2* that increases fasting glucose levels increases between-subject variance (5.7% per allele, $P = 4.2 \times 10^{-10}$), whereas variants in *GCK* and *G6PC2* that increase fasting glucose levels decrease between-subject variance (7.5% per allele, $P = 4.9 \times 10^{-11}$ and 7.3% per allele, $P = 7.5 \times 10^{-18}$, respectively). Variants that increase mean and between-subject variance in fasting glucose levels tend to increase T2D risk, whereas those that increase the mean but reduce variance do not ($r^2 = 0.61$). The variants that increase between-subject variance increase fasting glucose heritability estimates. Intuitively, our results show that increasing the mean and variance of glucose levels is more likely to cause pathologically high glucose levels than increase in the mean offset by a decrease in variance.

Despite recent advances in the genetics of T2D, understanding of the pathophysiology of the disease is still limited. Genome-wide association studies have yielded over 80 variants that associate with T2D, fasting glucose levels and other glycemic traits[2–6]. Although there is overlap between loci that affect fasting glucose and those that affect T2D, the effects of variants on mean fasting glucose do not predict their effects on T2D[1]. Further, none of the eight variants that associate with hemoglobin A1c (HbA1c), but not fasting glucose, associate with T2D, although HbA1c values above 6.5% are used as a diagnostic criterion for T2D[1].

Most reports on analysis of loci associated with quantitative traits have been confined to the effects of variants on the means of traits. However, variants can also affect the variability of traits (variance heterogeneity)[7]. Such loci have been reported for some human traits, including the major histocompatibility complex (MHC) region for rheumatoid arthritis[8], *FTO* for body mass index (BMI)[9], *SLC2A9* for serum urate[10], *LEPR* for C-reactive protein and *ICAM1* for soluble ICAM1 (ref. 11), as well as for traits in other species like rats[12], flies[13] and plants[14]. Further, variants can also affect the variability in measurements taken from the same individual. We refer to these two types of variability as between-subject and within-subject variance. Here we estimate the variance effects of variants that have been associated with fasting glucose levels[1] and examine how their effects on variance correlate with their effects on T2D risk. We also estimate how the effects of these variants on variance affect heritability estimates.

We chip genotyped 117,548 Icelanders with glucose measurements performed at three laboratories (**Fig. 1**, **Table 1**, **Supplementary Fig. 1** and **Supplementary Tables 1–4**). Of the subjects, 8,797 (7.5%) had T2D or were on diabetes medication[15]. Furthermore, 366 individuals had type 1 diabetes (T1D). The primary glucose variance association analysis was performed on individuals with fasting glucose levels (set I). Additionally, we generated three data sets for secondary analysis; one comprising individuals with fasting and/or non-fasting glucose levels (set II) and the previously listed data sets I and II after excluding T2D and T1D cases and individuals on diabetes medication.

Of the 36 known variants associated with glucose levels[1], 3 associated with between-subject variance consistently in all four analyses ($P < 0.05/36 = 0.0014$) (**Fig. 1a** and **Supplementary Tables 3–5**). One variant, rs7903146 in *TCF7L2*, is the strongest common T2D-associated variant[2,16]. The allele at this SNP associating with higher glucose levels and increased T2D risk was associated with greater between-subject glucose variance. In contrast, the alleles of rs560887 in *G6PC2* and rs2908289 in *GCK* that are associated with increased

[1]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [4]Department of Endocrinology and Metabolic Medicine, Landspitali, National University Hospital of Iceland, Reykjavik, Iceland. [5]Children's Medical Center, Landspitali, National University Hospital of Iceland, Reykjavik, Iceland. [6]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. [7]Laboratory in Mjodd, RAM, Reykjavik, Iceland. [8]Department of Clinical Biochemistry, Akureyri Hospital, Akureyri, Iceland. [9]Department of Clinical Biochemistry, Landspitali, National University Hospital of Iceland, Reykjavik, Iceland. [10]Iranian Molecular Medicine Network, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran. [11]Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Correspondence should be addressed to D.F.G. (daniel.gudbjartsson@decode.is) or K.S. (kari.stefansson@decode.is).

Received 30 March; accepted 11 July; published online 7 August 2017; doi:10.1038/ng.3928

Table 1 Summary of the data

						n measurements		T2D		T1D		Age		YOB	
	n	Mean	Q1	Median	Q3	Mean	Range	*n*	%	*n*	%	Mean	s.d.	Mean	s.d.
Fasting glucose levels (set I)															
Male	28,981	5.8	5.0	5.5	6.2	3.0	1–55	3,296	11.4	76	0.3	61.6	15.2	1947.4	16.1
Female	40,161	5.4	4.8	5.2	5.7	3.0	1–94	3,059	7.6	78	0.2	59.0	17.1	1950.5	18.1
All glucose levels (set II)															
Male	51,911	6	5	5.6	6.5	8.2	1–234	4,676	9.0	185	0.4	62.9	16.2	1943.9	16.7
Female	65,637	5.6	4.8	5.3	6.0	7.8	1–280	4,121	7.3	181	0.3	60.3	18.2	1946.6	18.7
HbA1c (first measurement)															
Male	18,107	5.8	5.2	5.5	5.9	–	–	3,041	16.8	56	0.3	60.1	15.1	1948.9	15.8
Female	22,945	5.6	5.2	5.5	5.8	–	–	2,676	11.7	47	0.2	56.9	17.2	1952.1	17.7

T2D, type 2 diabetes; T1D, type 1 diabetes; YOB, year of birth; Q1, first quartile; Q3, third quartile.

glucose[1] associated with less between-subject variance. The variant in *G6PC2* does not associate significantly with T2D whereas the variant in *GCK* slightly increases T2D risk in the DIAGRAM Consortium (odds ratio (OR) = 1.04, $P = 0.018$; **Supplementary Table 2**).

We also estimated the effects of the 36 variants on the within-subject variance in glucose levels (**Fig. 1b**). The glucose-increasing alleles of three variants—rs560887 in *G6PC2*, rs6943153 in *GRB10* and rs2908289 in *GCK*—associated consistently with less within-subject variance in all four analyses (**Supplementary Tables 3–5**).

On the basis of a T2D meta-analysis (12,171 cases and 56,862 controls of European ancestry)[2], 22 of the 36 variants with an effect on mean fasting glucose levels also associate with T2D. However, their effects on fasting glucose levels and T2D risk were weakly correlated ($r^2 = 0.02$ between the effect on the mean (β) and log(OR), $P = 0.21$; an F test was performed in all regression analysis) (**Fig. 2a**). Interestingly, the effect of a variant on between-subject variance in fasting glucose combined with its effect on mean fasting glucose predicted the effect of this variant on T2D much better than the effect on the mean alone ($r^2 = 0.61$, P value for adding effect on between-subject variance = 5.7×10^{-8}). Even on its own, the effect on between-subject variance predicted the T2D effect reasonably well ($r^2 = 0.38$, $P = 3.3 \times 10^{-5}$) (**Fig. 2b**). Therefore, variants that increase both the mean and between-subject variance of glucose levels increase the risk of T2D more than variants that increase the mean but reduce the between-subject variance.

The effect on within-subject glucose variance was a worse predictor of T2D risk than the effect on between-subject variance ($r^2 = 0.24$) (**Supplementary Table 6**), and it did not improve prediction of T2D beyond the mean and between-subject effects ($P = 0.091$).

Interaction between sequence variants and environmental factors such as nutrition is a possible source of between-subject variance[11]. It has previously been reported that heterogeneity in T2D associations is introduced by BMI[17,18]. We estimated the interaction effects between the 36 glucose-associated variants and BMI on fasting glucose (n = 39,986). The interaction effects were correlated with the between-subject variance effects ($r^2 = 0.12$, $P = 0.020$) (**Supplementary Fig. 2** and **Supplementary Table 7**). These results show that the effects of variants are affected by environment, although only a small fraction of the effects on between-subject variance are mitigated through interaction with BMI.

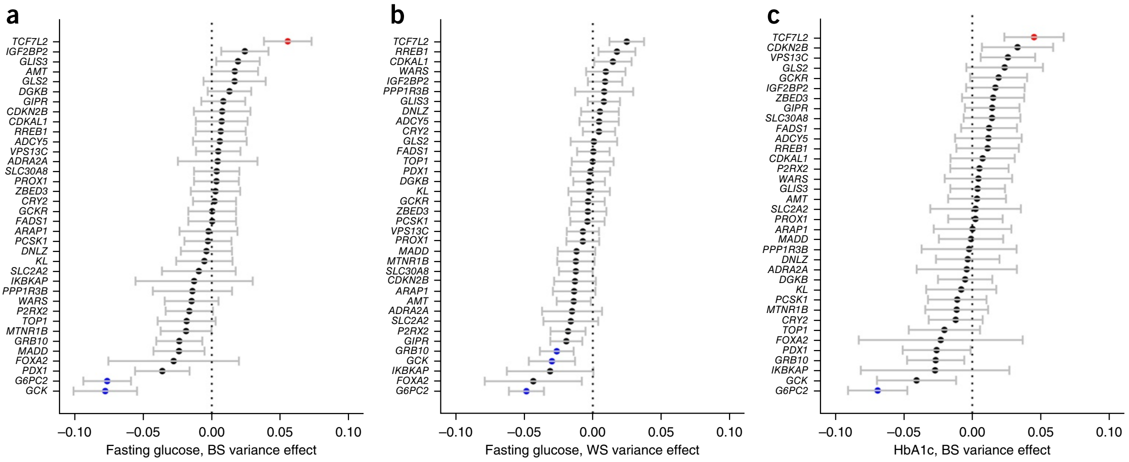

Figure 1 Effects of 36 published fasting-glucose-associated variants on between-subject and within-subject variance in fasting glucose levels and between-subject variance in HbA1c levels. Effects on variance are given for the allele that increases fasting glucose levels (**Supplementary Table 3**). Variants are colored blue if they significantly decrease the variance and red if they significantly increase it (likelihood-ratio test, $P < 0.05/36 = 0.0014$). (**a**) Effects on between-subject variance in fasting glucose ($\log(\alpha_{BS})$) and 95% confidence intervals for the estimated effects. (**b**) Effects on within-subject variance in fasting glucose levels ($\log(\alpha_{WS})$) and 95% confidence intervals for the estimated effects. (**c**) Effects on between-subject variance in HbA1c ($\log(\alpha_{BS})$) and 95% confidence intervals for the estimated effects.

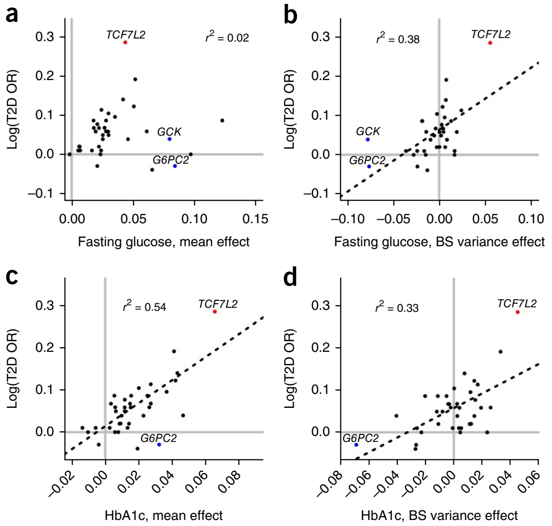

Figure 2 Effects of 36 published fasting-glucose-associated variants on fasting glucose and HbA1c, and between-subject variance in fasting glucose and HbA1c versus their effects on type 2 diabetes risk. Effects on fasting glucose were estimated in the Icelandic data, while effects on T2D risk were obtained from a T2D meta-analysis[2] (T2D-GENES Consortium, GoT2D Consortium, DIAGRAM Consortium; see URLs) (**Supplementary Tables 2**, **3** and **11**). Effects are given for the allele that increases fasting glucose levels. Variants are colored blue if they significantly decrease variance and red if they significantly increase it ($P < 0.05/36 = 0.0014$). (**a**) Fasting glucose mean effect (β) against log(T2D OR). (**b**) Fasting glucose between-subject variance effect ($\log(\alpha_{BS})$) against log(T2D OR). (**c**) HbA1c mean effect (β) against log(T2D OR). (**d**) HbA1c between-subject variance effect ($\log(\alpha_{BS})$) against log(T2D OR).

An undetected secondary variant can create a variance effect for the primary variant. However, secondary signals at the loci associated with between-subject variance had no impact on variance effects (**Supplementary Table 8**). Another possible source of effects on between-subject variance is interaction between loci. For the three variants associated with between-subject variance, we found no interaction (**Supplementary Table 9**).

To validate these variance effects, we analyzed a sample of 10,437 Iranians with 44,470 fasting glucose measurements from the prospective Tehran Lipid and Glucose Study[19]. We replicated the association of the variants in *TCF7L2*, *GCK* and *G6PC2* with between-subject variance and the association of the *G6PC2* and *GCK* variants with within-subject variance (**Supplementary Tables 5** and **10**).

HbA1c reflects the average plasma glucose concentration over 3 months, and an HbA1c value above 6.5% is used as a diagnostic criterion for T2D[15]. HbA1c measurements were available for 41,052 Icelanders with genotype information (**Table 1** and **Supplementary Table 1**). The number of measurements per subject was correlated with HbA1c. Therefore, we only used the first measurement for each subject in our analysis.

The pattern of effect for the 36 markers on between-subject HbA1c variability is consistent with the results for fasting glucose (**Fig. 1**, **Supplementary Fig. 3** and **Supplementary Table 11**). Of the 36 variants, the variants in *TCF7L2* and *G6PC2* were associated with between-subject variance (4.5% increase per allele, $P = 4.5 \times 10^{-5}$ and 6.9% decrease per allele, $P = 4.0 \times 10^{-10}$, respectively; a likelihood-ratio test was performed in all genome-wide associations). As for fasting glucose, the effect on between-subject variance in HbA1c increased the prediction accuracy of the effect on T2D ($r^2 = 0.54$ for the mean only, $r^2 = 0.77$ for the mean and between-subject variance effect, P value for adding the between-subject variance effect $= 1.4 \times 10^{-6}$) (**Fig. 2c**,**d** and **Supplementary Table 6e**).

Eight variants have been reported to affect HbA1c without affecting fasting glucose, none of which have an effect on T2D[1,20]. These variants associate with red blood cell homeostasis and iron metabolism (**Supplementary Table 12**). Interestingly, the HbA1c-increasing allele for all eight markers lowered between-subject variance (**Supplementary Figs. 4** and **5**, and **Supplementary Table 13**), of which two were significantly associated with lower between-subject variance ($P < 0.05/8 = 0.0063$): rs10159477[G] in *HK1* was associated with 5.1% lower variance per allele ($P = 0.0024$) and rs6474359[T] in *ANK1* was associated with 8.0% lower variance per allele ($P = 0.0044$). The increase in the mean was offset by lower variance for carriers of these variants, and these individuals are therefore less likely to have high HbA1c measures. This may explain why carriers of these HbA1c-increasing variants are not likely to be misclassified as diabetic[20].

We constructed genetic risk scores (GRSs), based on the 36 variants, for both mean and between-subject variance of fasting glucose levels. Both GRSs were associated with T2D ($P < 3.1 \times 10^{-39}$; **Fig. 3** and **Supplementary Table 14**). Adding the GRS for between-subject variance to the GRS for the mean increased residual Nagelkerke's pseudo-r^2 from 0.4% to 1.0% ($P = 5.4 \times 10^{-67}$; **Supplementary Table 14**). Similarly, GRSs based on the 36 variants for glucose levels and the 8 variants for HbA1c measures were associated with T2D ($P < 3.4 \times 10^{-28}$; **Fig. 3** and **Supplementary Table 14**). This shows that the effects of variants on between-subject variance have an impact on genetic T2D risk prediction that is comparable to that from their effects on the mean.

The heritability of a trait is the fraction of variance attributable to genetics. Classical estimates of heritability ignore the impact of variants on phenotypic variance. Most heritability estimates are based on relating the correlation between relative pairs to the genetic sharing between relatives[21]. Correlation between relatives corresponds to the ratio of their covariance and the geometric mean of their phenotypic variances. Variants that affect variance will have a substantial impact on the denominator. However, their effect on covariance is unpredictable. In our data, we had fasting glucose measures and genotypic information for 35,965 sibling pairs and 38,527 parent–offspring pairs. To investigate the effect of variants on the covariance between relatives, we calculated the covariance for genotype-concordant relative pairs and estimated the relationship between genotype and covariance. For the 36 variants associated with glucose levels, the mean covariance trend in siblings and parent–offspring pairs correlated positively with the between-subject variance effect ($r^2 = 0.22$, $P = 2.1 \times 10^{-3}$) (**Fig. 4a** and **Supplementary Table 15**). If the increase in covariance per allele was higher than the variance effect, the correlation was also increased and the variants therefore also increased the estimated narrow-sense heritability. The variant in *TCF7L2* had the strongest trend of 17.6% increased covariance ($P = 4.1 \times 10^{-4}$) (**Fig. 4b**). The between-subject variance effect of *TCF7L2* was 5.7% per allele, and the correlation was therefore increased by 11.3% per allele.

We have shown that variants in *TCF7L2*, *GCK*, *G6PC2* and *GRB10* that affect mean fasting glucose levels also associate with variance in glucose. The variance effects remain after the removal of diabetic cases and individuals on diabetes medication. The two variants that lower between-subject variance do not associate with T2D risk, and their variance effect is thus not driven by a diabetes medication.

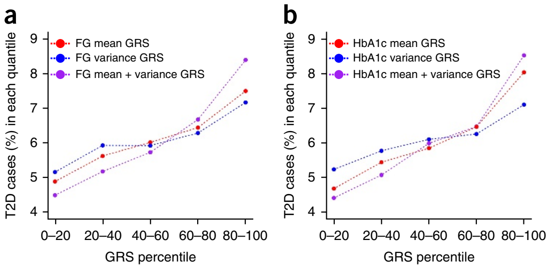

Figure 3 The percentage of type 2 diabetes cases in each quantile of the genetic risk scores. Combination of the GRSs for the mean and between-subject variance was weighted with the coefficients from logistic regression between T2D and the GRSs (**Supplementary Table 15**). (**a**) Fasting glucose (FG) GRS based on the 36 fasting-glucose-associated variants. (**b**) HbA1c GRS based on the 36 fasting-glucose-associated variants and 8 HbA1c-associated variants.

Conversely, removal of diabetic cases could create a variance effect in the presence of an effect on the mean, although we do not observe this phenomenon in our data. It is, however, likely that variants' effects on variance are at least partly due to their interaction with other variants and/or with environmental factors. This hypothesis is supported by the correlation between the variants' between-subject variance effects and their interaction with BMI.

We have also shown that variants that increase both mean fasting glucose levels and between-subject glucose variance increase T2D risk more than variants that increase fasting glucose but reduce the between-subject variance. These results largely account for the apparent discrepancy between the effects of variants on fasting glucose and their effects on T2D risk. This result is intuitively appealing, as T2D is primarily a disease of too high glucose; variants that increase both the mean and variance for glucose are more likely to be associated with pathologically high glucose levels than variants that only increase the mean or even have an increase in the mean offset by lower variance.

The variants in *GCK*, *G6PC2* and *TCF7L2* all affect fasting glucose levels, but their effects on T2D risk are not proportionate to their effects on glucose[22]. This may reflect different roles in glucose regulation. *GCK* and *G6PC2* encode enzymes that regulate glucose homeostasis, effectively establishing the glucose set point. Variants that increase mean glucose through these proteins will be countered by pressure to keep the glucose level within the physiological range, leading to reduced variance associated with these variants both within and between subjects. Similarly, variants that associate with increased HbA1c but not fasting glucose or T2D all associate with erythrocyte physiology and iron homeostasis and, where significant, lower HbA1c variance. Overall, this indicates low tolerance for variability in homeostatic regulation. In contrast, the variant associated with the highest variance in glucose levels is located in *TCF7L2*, which encodes a transcription factor that is thought to affect glucose levels through complex regulation of beta cell mass and function[23]. This variant affects beta cell response to glucose, leading to greater sensitivity to the environment and, thus, greater variability in glucose levels among carriers.

Only 2% of the heritability of fasting glucose levels is attributable to the effect of the 36 glucose-associated variants on mean levels. We have shown that variants that increase between-subject variance create positive covariance between individuals beyond their effects on the mean, increasing heritability estimates based on correlation between relative pairs. The effect of these markers on heritability is substantial and so is their contribution to the missing heritability of fasting glucose levels. Further, the effects of variants on the variability between individuals in glucose and HbA1c levels are as important for genetic risk prediction as the effects of variants on the mean.

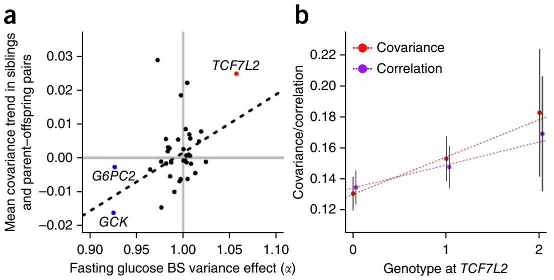

Figure 4 Covariance between genotype-concordant relative pairs. (**a**) Effects of 36 published fasting-glucose-associated variants on between-subject variance in fasting glucose levels and their glucose level covariance trends in pairs of relatives (**Supplementary Table 15**). Effects are given for the allele that increases fasting glucose levels. Variants are colored blue if they significantly decrease the variance and red if they significantly increase it ($P < 0.05/36 = 0.0014$). (**b**) Estimated covariance and correlation of fasting glucose measurements among pairs of relatives with the same genotype at *TCF7L2* and the 95% confidence intervals for the covariance and correlation estimates.

URLs. T2D-GENES Consortium, GoT2D Consortium, DIAGRAM Consortium (2016-09-12), http://www.type2diabetesgenetics.org/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank the subjects of the Icelandic deCODE study and the Iranian study for their participation. We also thank the staff at deCODE Genetics core facilities and all our colleagues for their contributions to this work. This research project has been supported by grant no. 121 NRCI Research Project and with the support of the National Research Council of the Islamic Republic of Iran.

AUTHOR CONTRIBUTIONS

E.V.I., V.S., P.S., H.H., U.T., D.F.G. and K.S. designed the study and interpreted the results. E.V.I. and D.F.G. performed statistical analysis. M.S.D., G.T., S.S., A.B.H., G.S., R. Bjarnason, A.V.T., R. Benediktsson, G.E., O.S., I.O., S.Z. and F.A. performed recruitment and phenotyping. The manuscript was drafted by E.V.I., V.S., D.F.G. and K.S. All authors contributed to the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
2. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
3. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
4. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).

5. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
6. Gaulton, K.J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
7. Rönnegård, L. & Valdar, W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genet.* **13**, 63 (2012).
8. Wei, W.-H. *et al.* Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. *Sci. Rep.* **6**, 25014 (2016).
9. Yang, J. *et al. FTO* genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).
10. Topless, R.K. *et al.* Association of *SLC2A9* genotype with phenotypic variability of serum urate in pre-menopausal women. *Front. Genet.* **6**, 313 (2015).
11. Paré, G., Cook, N.R., Ridker, P.M. & Chasman, D.I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genet.* **6**, e1000981 (2010).
12. Perry, G.M.L. *et al.* Sex modifies genetic effects on residual variance in urinary calcium excretion in rat (*Rattus norvegicus*). *Genetics* **191**, 1003–1013 (2012).
13. Mackay, T.F. & Lyman, R.F. *Drosophila* bristles and the nature of quantitative genetic variation. *Phil. Trans. R. Soc. Lond. B* **360**, 1513–1527 (2005).
14. Shen, X., Pettersson, M., Rönnegård, L. & Carlborg, Ö. Inheritance beyond plain heritability: variance-controlling genes in *Arabidopsisthaliana*. *PLoS Genet.* **8**, e1002839 (2012).
15. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
16. Grant, S.F. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
17. Perry, J.R. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in *LAMA1* and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.* **8**, e1002741 (2012).
18. Cauchi, S. *et al.* The genetic susceptibility to type 2 diabetes may be modulated by obesity status: implications for association studies. *BMC Med. Genet.* **9**, 45 (2008).
19. Azizi, F. *et al.* Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials* **10**, 5 (2009).
20. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A_{1C} levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229–3239 (2010).
21. Falconer, D.S. & Mackay, T.F.C. in *Introduction to Quantitative Genetics* 4th edn, Ch. 10 (Pearson, 1996).
22. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
23. Mitchell, R.K. *et al.* Selective disruption of *Tcf7l2* in the pancreatic β cell impairs secretory function and lowers β cell mass. *Hum. Mol. Genet.* **24**, 1390–1399 (2015).

ONLINE METHODS

Study subjects. *Iceland.* Measurements of glucose levels were available for a total of 117,548 Icelanders genotyped using Illumina chips. All study participants provided informed consent, and the study was approved by the Data Protection Commission of Iceland and the Icelandic National Bioethics Committee.

Iran. The Iranian subjects are part of the ongoing Tehran Lipid and Glucose Study[19], including 10,437 Iranians with 44,470 fasting glucose measurements genotyped using Illumina chips. All study participants provided informed consent. The study has been approved by the National Research Council of the Islamic Republic of Iran (no. 121) and has been performed with the approval of the Human Research Review Committee of the Endocrine Research Center, Shahid Beheshti University (M.C.).

SNP selection. The 36 fasting-glucose-associated variants were identified in a genome-wide association meta-analysis of up to 133,010 individuals of European ancestry without diabetes, including individuals genotyped using the Metabochip[1].

Whole-genome sequencing. The process used for whole-genome sequencing of the 8,453 Icelanders and the subsequent imputation have been described in a recent publication[24].

Association testing. *Mean effect.* Both fasting and non-fasting glucose measurements were transformed to a standard normal distribution using a rank-based inverse-normal transformation within each sex and each source separately and adjusted for age at measurement using a generalized additive model[25]. For each SNP, a classical linear regression, using the genotype as an additive covariate and mean glucose levels per subject as a response, was fit to test for association.

Between-subject variance effect. For each SNP, we fit a normal model where the mean glucose level per subject was regressed against the genotype and the between-subject variance was assumed to change multiplicatively with the genotype so that for non-carriers, heterozygotes and homozygotes the between-subject variance was assumed to be σ^2, $\alpha_{BS}\sigma^2$ and $\alpha_{BS}^2\sigma^2$, respectively (**Supplementary Note**).

Within-subject variance effect. For each SNP, we fit a normal model where glucose level measurements were regressed against the genotype and the within-subject variance was assumed to change multiplicatively with the genotype so that for non-carriers, heterozygotes and homozygotes the within-subject variance was assumed to be σ^2, $\alpha_{WS}\sigma^2$ and $\alpha_{WS}^2\sigma^2$, respectively (**Supplementary Note**).

Subjects in the data sets were related, causing the χ^2 test statistic to have mean >1 and median >0.675. We used a subset of 640,250 common SNPs to estimate the inflation factor λ and computed all P values by dividing the corresponding χ^2 values by λ to adjust for both relatedness and potential population stratification[26]. For the fasting glucose data set (I), $\lambda = 1.14$, and $\lambda = 1.21$ when estimating between-subject and within-subject variance effects, respectively.

BMI interaction effect. For each SNP, we fit an interaction regression model, using the genotype, BMI and the interaction term between the genotype and BMI as covariates and mean fasting glucose levels as the response. Both glucose levels and BMI measurements were transformed to a standard normal distribution using a rank-based inverse-normal transformation within each sex and each source separately and adjusted for age at measurement using a generalized additive model[25].

Thresholds for significance. In the set of 36 variants, significance thresholds for between-subject and within-subject variance effect were set to control the false discovery rate at 5% using standard Bonferroni correction ($P < 0.05/36 = 0.0014$).

Trend analysis. We assessed the relationship between the effects of sequence variants on mean and variance effects on glucose levels and their effect on T2D (log(OR)) using the following models:

A. T2D effect versus glucose mean effect: $\log(\mathrm{OR}) = y_1\beta + \varepsilon$;
B. T2D effect versus glucose between-subject variance effect: $\log(\mathrm{OR}) = y_2\log(\alpha_{BS}) + \varepsilon$;
C. T2D effect versus glucose mean and between-subject variance effect: $\log(\mathrm{OR}) = y_1\beta + y_2\log(\alpha_{BS}) + \varepsilon$;
D. T2D effect versus glucose mean effect, between-subject variance effect and the interaction between glucose mean and between-subject variance effect: $\log(\mathrm{OR}) = y_1\beta + y_2\log(\alpha_{BS}) + y_3(\beta \times \log(\alpha_{BS})) + \varepsilon$;
E. T2D effect versus glucose within-subject variance effect: $\log(\mathrm{OR}) = y_4\log(\alpha_{WS}) + \varepsilon$;
F. T2D effect versus glucose mean and within-subject variance effect: $\log(\mathrm{OR}) = y_1\beta + y_4\log(\alpha_{WS}) + \varepsilon$;
G. T2D effect versus glucose mean, between-subject variance and within-subject variance effect: $\log(\mathrm{OR}) = y_1\beta + y_2\log(\alpha_{BS}) + y_4\log(\alpha_{WS}) + \varepsilon$

where β is the glucose mean effect, α_{BS} is the between-subject variance effect and α_{WS} is the within-subject variance effect. All models were fitted with a simple weighted linear regression where each variant was weighted by $f(1-f)$, where f is the minor allele frequency of the variant, such that rare variants have less weight in the computation than common variants. The estimates and measures of goodness of fit are given in **Supplementary Table 6**.

Genetic risk scores. GRSs were constructed for both fasting glucose and HbA1c levels by combining the effect allele counts for the selected variants weighted by either the estimated mean effect or the between-subject variance effect of each allele on the trait.

Heritability. The correlation between close relative pairs is usually used to estimate heritability[21]. To assess how much variants effecting between-subject variance can contribute to heritability estimates, for each SNP, we estimated the covariance between siblings having the same genotype. Then, we performed a weighted linear regression between the estimated covariance and the genotype to assess the covariance trend. We weighted by the number of siblings having the genotype divided by the squared phenotypic variance given the genotype (**Supplementary Note**). This was repeated for parent–offspring pairs. The correlation between relatives is the ratio of their covariances and the geometric mean of their phenotypic variances. The correlation trend was therefore computed as the ratio of the covariance trend and variance trend (**Supplementary Note**).

A **Life Sciences Reporting Summary** for this paper is available.

Code availability. The code used to detect between-subject and within-subject variance effects is available as **Supplementary Code**.

Data availability. The authors declare that the data supporting the findings of this study are available within the article, its supplementary information files and upon request.

24. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
25. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310 (1986).
26. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).

Paper II

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09304-9 OPEN

Sequence variation at *ANAPC1* accounts for 24% of the variability in corneal endothelial cell density

Erna V. Ivarsdottir[1,2], Stefania Benonisdottir[1], Gudmar Thorleifsson[1], Patrick Sulem[1], Asmundur Oddsson[1], Unnur Styrkarsdottir[1], Snaedis Kristmundsdottir[1], Gudny A. Arnadottir[1], Gudmundur Thorgeirsson[1,3,4], Ingileif Jonsdottir[1,3,5], Gunnar M. Zoega[6], Unnur Thorsteinsdottir[1,3], Daniel F. Gudbjartsson[1,2], Fridbert Jonasson[3,6], Hilma Holm[1] & Kari Stefansson[1,3]

The corneal endothelium is vital for transparency and proper hydration of the cornea. Here, we conduct a genome-wide association study of corneal endothelial cell density (cells/mm^2), coefficient of cell size variation (CV), percentage of hexagonal cells (HEX) and central corneal thickness (CCT) in 6,125 Icelanders and find associations at 10 loci, including 7 novel. We assess the effects of these variants on various ocular biomechanics such as corneal hysteresis (CH), as well as eye diseases such as glaucoma and corneal dystrophies. Most notably, an intergenic variant close to *ANAPC1* (rs78658973[A], frequency = 28.3%) strongly associates with decreased cell density and accounts for 24% of the population variance in cell density ($\beta = -0.77$ SD, $P = 1.8 \times 10^{-314}$) and associates with increased CH ($\beta = 0.19$ SD, $P = 2.6 \times 10^{-19}$) without affecting risk of corneal diseases and glaucoma. Our findings indicate that despite correlations between cell density and eye diseases, low cell density does not increase the risk of disease.

[1] deCODE genetics/Amgen, Reykjavik, Iceland. [2] School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3] Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. [4] Division of Cardiology, Department of Internal Medicine, Landspitali, The National University Hospital of Iceland, Reykjavik, Iceland. [5] Department of Immunology, Landspitali, The National University Hospital of Iceland, Reykjavik, Iceland. [6] Department of Ophthalmology, Landspitali, The National University Hospital of Iceland, Reykjavik, Iceland. Correspondence and requests for materials should be addressed to H.H. (email: hilma.holm@decode.is) or to K.S. (email: kari.stefansson@decode.is)

Corneal diseases are among the most common causes of visual loss worldwide and endothelial cell failure is the leading indication for corneal transplantation[1]. The corneal endothelium is the monolayer of cells at the innermost surface of the cornea. Through an ion pump function, the endothelium is responsible for balanced corneal hydration, thus maintaining transparency by preventing edema and disruption of lamellar spacing of the collagen fibrils in the corneal stroma[2]. Maintaining proper function of the endothelium requires a minimum number of endothelial cells, around 400–500 cells/mm^2 [3]. The cells are generally thought to be incapable of mitosis after birth but are halted in the G1 phase of the cell cycle and their number decreases with age[2]. The cell division is thought to be blocked by contact inhibition, high concentration of negative growth factors in the anterior chamber and by accumulation of reactive oxygen species promoting state of stress-induced senescence[3,4]. The response to cell loss includes spreading and/or migration of adjacent cells which increase in size and become more variable in both cell size and shape[3,5].

Non-contact auto-tracking and focusing specular microscopy provides a non-invasive analysis of the structure of the corneal endothelium[6]. The equipment captures an image of the endothelial cell layer and provides measures of its structure including, cell density (cells/mm^2), coefficient of cell size variation (CV), percentage of hexagonally shaped cells (HEX) and central corneal thickness (CCT). CCT has been studied extensively in large genome-wide association studies (GWAS) where the measurements are obtained using different types of instruments[7–13]. Several CCT associating loci have been identified including *WNT10A, COL5A1*, and *ZNF469*[11]. To our knowledge, however, there are no reports of sequence variants influencing other direct measures of endothelial structure such as cell density, CV and HEX.

The measures of endothelial structure are used to diagnose corneal diseases. Several corneal diseases including Fuchs endothelial corneal dystrophy (FECD) and macular corneal dystrophy (MCD) are known to have genetic components[14,15]. FECD is a leading cause of corneal transplant surgery and is characterized by premature loss of endothelial cells resulting in increased variability in cell shape and size leading to corneal edema and visual loss[16,17]. MCD is characterized by progressive spotted corneal opacities leading to severe visual impairment, caused by homozygous or compound heterozygous variants in the *CHST6* gene (OMIM: 605294). MCD is a rare condition worldwide, but because of the founder effect[18], MCD is unusually common in Iceland where it accounts for approximately one-third of corneal transplantations[19].

Cell density in the corneal endothelium may be reduced in glaucoma patients[20]. Glaucoma is an ocular disease that affects ~3.5% of people over 40 years of age, and is a major cause of irreversible blindness worldwide[21,22]. Commonly, elevated intraocular pressure (IOP) leads to progressive damage of the optic nerve causing visual loss and it has been postulated that elevated IOP affects the endothelium[23]. The relationship between different corneal measures and glaucoma have been investigated, especially the role of IOP, CCT and more recently corneal hysteresis (CH) and corneal resistance factor (CRF)[24]. CH and CRF are measures of corneal response to a rapid jet of air, where CH is a measure of the elasticity of the cornea and CRF is an overall indicator of the resistance of the cornea[25]. Lower CH has been associated with faster rate of glaucoma progression[26–28].

Here we describe our search for sequence variants associating with measures of corneal structure, finding 10 sequence variants, nine common and one low-frequency (minor allele frequency (MAF) <5%), associating with CCT, cell density, CV, or HEX. Two of these variants satisfy thresholds for genome-wide significance for more than one trait. Seven of the associations are novel, two are represented by coding variants and one is located in a gene known to cause a Mendelian disorder (*ADAMTS17*, OMIM: 607511). We assess the effects on ocular biomechanics, including IOP and CH, and various eye diseases. We find a variant near *ANAPC1* that strongly associates with cell density but not with risk of eye disease, which indicates that low cell density alone does not affect disease development directly.

Results

Summary of the data. Endothelial images from a specular microscopy of 6125 Icelanders were used in the analysis, providing measures of cell density, CV, and HEX. The equipment also produced CCT measurements. These images were obtained as a part of a comprehensive phenotyping of a general population sample (the deCODE health study), currently including 6300 Icelanders that were between 18 and 94 years of age at the time of recruitment (44.6% men; mean age = 55.9, standard deviation (SD) = 15.1). We also measured several ocular biomechanics such as CH, CRF, Goldmann correlated intraocular pressure (IOPg), and corneal compensated intraocular pressure (IOPcc) (Supplementary Figures 1–6).

The number of endothelial cells declines with age and remaining cells enlarge to compensate for the cell loss. Consequently, cell density and HEX decrease with age while CV increases (Table 1, Fig. 1a–c and Supplementary Figure 7). Women have considerably lower HEX than men (50.6% vs 48.3%, $P = 8.4 \times 10^{-43}$; F test), while cell density and CV are slightly higher for women (2663 vs 2639 cells/mm^2, $P = 1.8 \times 10^{-3}$ (F test) and 30.3 vs 29.6, $P = 2.6 \times 10^{-6}$ (F test), respectively). To our knowledge, the gender differences of HEX, cell density and

Table 1 Summary of data

	Mean (SD)	Men's mean (SD)	Women's mean (SD)	Sex effect	Sex *P*-value	Glaucoma effect [SD]	Glaucoma *P*-value
CD	2652 (296)	2639 (296)	2663 (294)	23.83	1.8×10^{-3}	−0.35	2.1×10^{-6}
CV	30.0 (5.9)	29.6 (6.7)	30.3 (5.1)	0.71	2.6×10^{-6}	0.13	0.069
CCT	563 (40)	565 (40)	562 (39.5)	−3.82	1.9×10^{-4}	−0.25	4.7×10^{-4}
HEX	49.3 (6.5)	50.6 (6.6)	48.3 (6.3)	−2.29	8.4×10^{-43}	−0.16	0.030
CH	10.4 (1.2)	10.2 (1.2)	10.5 (1.1)	0.28	5.7×10^{-22}	−0.37	3.8×10^{-7}
IOPg	14.7 (3.4)	14.6 (3.5)	14.7 (3.4)	0.13	0.15	0.18	0.014
IOPcc	15.3 (3.1)	15.4 (3.2)	15.2 (3.0)	−0.19	0.015	0.29	7.3×10^{-5}
CRF	13.1 (1.5)	13.0 (1.5)	13.3 (1.5)	0.29	5.7×10^{-14}	−0.07	0.34

The mean and standard deviation (SD) is shown for each corneal trait obtained from the specular microscopy equipment and the ocular response analyzer, overall and separately for each sex. The effect of sex and glaucoma status on each trait and the corresponding *P*-values (F test) are shown. The sample size was 6125 in total, 2733 men and 3392 women

CD cell density, *CV* coefficient of cell size variation, *CCT* central corneal thickness, *HEX* percentage of hexagonal cells, *CH* corneal hysteresis, *IOPg* Goldmann correlated intraocular pressure, *IOPcc* corneal compensated intraocular pressure, *CRF* corneal resistance factor

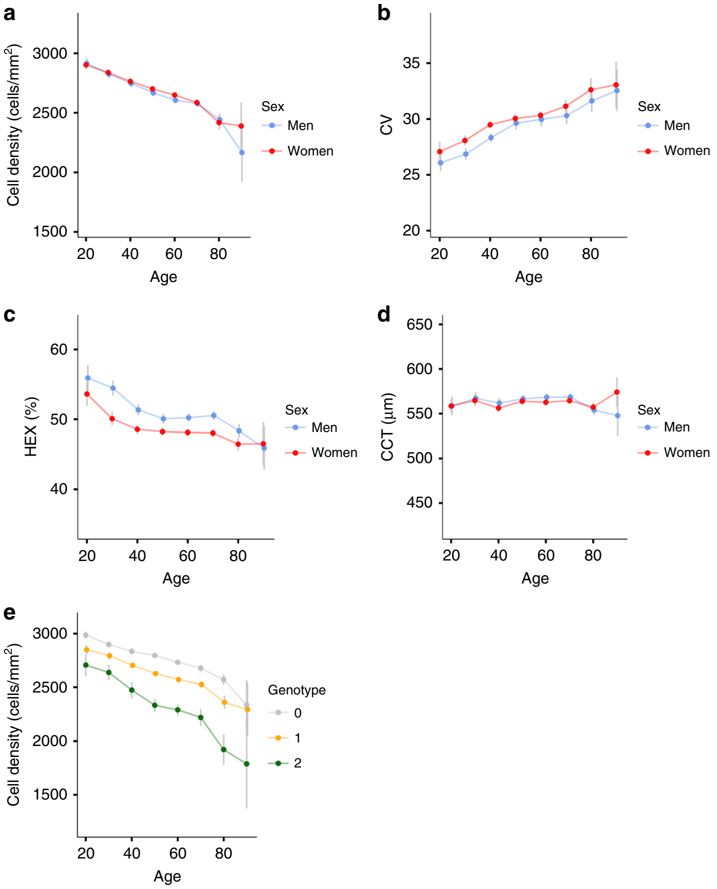

Fig. 1 Corneal structure measurements by age ($N = 6125$). The average **a** cell density, **b** HEX, **c** CV, and **d** CCT values for subjects belonging to a 10 year age group (e.g., age = 30 for individuals between 26 and 35) against age, for men and women. **e** The average cell density for subjects belonging to a 10 year age group against age for noncarriers, heterozygous, and homozygous carriers of the *ANAPC1* variant, rs78658973. The gray lines show the 95% confidence intervals

CV have not been reported before. Consistent with previous reports[29], CCT does not change with age and is higher for men than women (565.4 vs 561.5 µm, $P = 1.9 \times 10^{-4}$; F test) (Fig. 1d). The measurements of corneal structure are correlated, also after adjusting for sex and age (Supplementary Table 1). The strongest correlations are between CV and HEX ($r = -0.65$) and between CV and cell density ($r = -0.33$).

Study design. To search for sequence variants associating with corneal structure, we analyzed 35.2 million sequence variants identified through whole-genome sequencing of 28,075 Icelanders that were subsequently imputed into 155,250 chip-typed individuals, as well as their first- and second-degree relatives[30,31] (Supplementary Figures 8–11). Ten sequence variants, satisfied our genome-wide significance thresholds that are dependent on sequence variant annotation[32] (Table 2, Supplementary Table 2).

To determine whether the associating variants affect the risk of eye diseases we performed a meta-analysis of GWAS results from Iceland and the UK Biobank (Supplementary Table 3 and 4). We tested the variants for association with glaucoma (8432 cases and 641,353 controls) and the sub-categories: primary open-angle glaucoma (2296 cases and 705,937 controls), and primary angle closure glaucoma (777 cases and 637,017 controls). We also tested the variants for association with disorders of cornea (ICD10 code H18, 756 cases and 663,218 controls) and the sub-categories; corneal degeneration (ICD10 code H18.4, 199 cases and 684,021 controls), hereditary corneal dystrophies (ICD10 code H18.5, 330 cases and 683,652 controls) and keratoconus (ICD10 code H18.6, 127 cases and 659,503 controls). We applied a Bonferroni corrected *P*-value threshold based on testing ten corneal structure variants for association with seven phenotypes ($P < 0.05/(7{*}10) = 7.1 \times 10^{-4}$).

GWAS results. Two sequence variants associate with cell density (Fig. 2a, Table 2). The strongest association is represented by a common intergenic variant located 0.4 kb downstream of *ANAPC1*, rs78658973[A] (MAF = 28.3%), that associates with decreased cell density ($\beta = -0.77$ SD, $P = 1.8 \times 10^{-314}$; a likelihood-ratio test was performed in all genome-wide associations) (Fig. 1e, Supplementary Figure 12). rs78658973 is highly correlated with 113 variants ($r^2 > 0.8$) in the region, none of which is protein coding. The most highly correlated coding variants are two splice region variants in *ANAPC1*; rs201128688 and rs142711068 ($r^2 = 0.72$ and 0.73, respectively). The effect of rs78658973[A] conditioning on the two splice region

Table 2 Association results

Trait	Chr:Position	rs-name	Allele (min/maj)	MAF (%)	Gene/ [Locus]	Coding effect	LD class	*P*-value	*β* [SD] (95% CI)	Ref.
CD	2:111726948	rs78658973	(A/T)	28.3	*[ANAPC1]*	Intergenic	112	1.8×10^{-314}	−0.77 (−0.77,−0.77)	
CD	18:55586154	–	CTG repeat > 33	6.1	*TCF4*		3	1.4×10^{-20}	−0.41 (−0.49,−0.32)	35
CV	2:111726948	rs78658973	(A/T)	28.3	*[ANAPC1]*	Intergenic	112	2.8×10^{-28}	0.23 (0.19,0.27)	
CV	17:14650919	rs2323458	(A/G)	36.1	*17p12*	Intergenic	47	6.9×10^{-13}	0.14 (0.10, 0.18)	
CV	8:9943404	rs10094779	(G/A)	24.1	*8p23.1*	Intergenic	4	7.6×10^{-12}	0.15 (0.11, 0.19)	
CV	11:122029470	rs76561503	(C/T)	17.9	*11q24.1*	Intergenic	29	3.3×10^{-10}	0.16 (0.11,0.21)	
CCT	16:88302168	rs12719930	(G/A)	39.4	*[ZNF469]*	Intergenic	21	1.9×10^{-14}	−0.15 (−0.19,−0.11)	13
CCT	2:218890289	rs121908120	(A/T)	2.6	*WNT10A*	Missense	6	4.5×10^{-11}	−0.39 (−0.51,−0.28)	9
CCT	9:134545337	rs943423	(G/A)	27.2	*[COL5A1]*	Intergenic	0	6.1×10^{-11}	−0.14 (−0.19,−0.10)	13
CCT	15:100152748	rs72755233	(A/G)	13.8	*ADAMTS17*	Missense	0	1.3×10^{-10}	0.18 (0.12, 0.23)	
CCT	12:104015054	rs117801489	(C/T)	4.3	*GLT8D2*	Missense	2	3.9×10^{-10}	0.30 (0.20, 0.39)	
HEX	18:55586154	–	CTG repeat > 33	6.1	*TCF4*		3	5.9×10^{-18}	−0.37 (−0.45,−0.28)	35
HEX	2:111726948	rs78658973	(A/T)	28.3	*[ANAPC1]*	Intergenic	112	2.8×10^{-13}	−0.16 (−0.20,−0.11)	

The 10 variants identified in the GWAS on cell density (CD), CV, HEX, and CCT. Effects are shown for the minor allele. Minor allele frequency in the Icelandic population is presented. The LD class column shows the number of highly correlated variants ($r^2 > 0.8$). The imputation information for all these variants is > 0.99

CD cell density, *CV* coefficient of cell size variation, *CCT* central corneal thickness, *HEX* percentage of hexagonal cells, *MAF* minor allele frequency, *LD* linkage disequilibrium, *CI* confidence interval

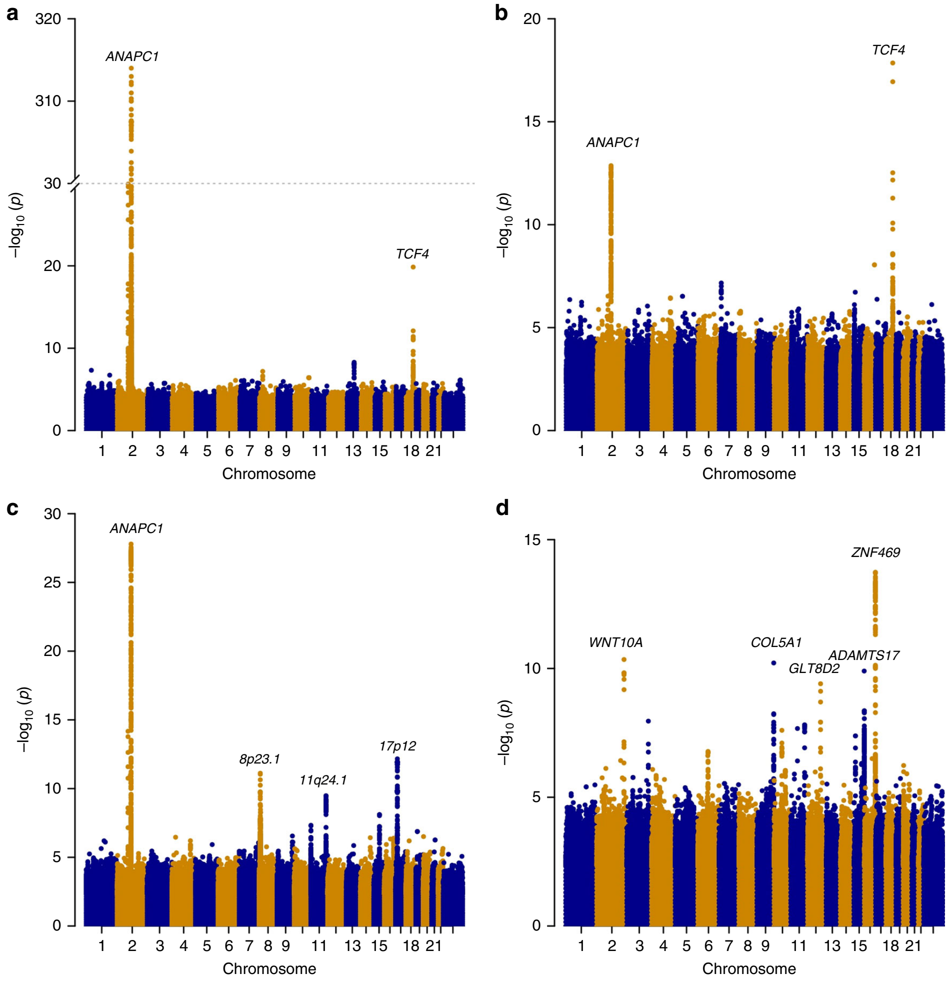

Fig. 2 Manhattan plots. Association results for the corneal measures ($N = 6125$) obtained from the specular microscopy; **a** cell density, **b** HEX, **c** CV, and **d** CCT. The $-\log_{10}$ *P*-values are plotted for each variant against their chromosomal position. A likelihood-ratio test was used when testing for association

variants is still significant ($\beta = -0.78$ SD, $P = 7.8 \times 10^{-85}$), while the effect of the splice region variants are completely explained by rs78658973. rs78658973[A] also associates with CV ($\beta = 0.23$ SD, $P = 2.8 \times 10^{-28}$) and HEX ($\beta = -0.16$ SD, $P = 2.6 \times 10^{-19}$), but this is largely driven by the strong effect on cell density (CV adjusted for cell density: $\beta = -0.04$ SD, $P = 0.049$; HEX adjusted for cell density: $\beta = -0.04$ SD, $P = 0.072$). Interestingly, rs78658973[A] also associates with CH ($\beta = 0.19$ SD, $P = 2.6 \times 10^{-19}$) (Supplementary Table 4, Fig. 3) which reflects the cornea's ability to absorb and dissipate energy[25,33]. CH and cell density are only weakly correlated after adjusting for sex and age ($r = -0.03$) and the effect of rs78658973[A] on CH is not affected by adjustment for cell density ($\beta_{\text{adjusted}} = 0.21$ SD, $P_{\text{adjusted}} = 1.5 \times 10^{-18}$). CH is lower in patients with glaucoma or corneal disorders like keratoconus[34]. However, rs78658973 does not associate with corneal diseases or glaucoma in our data ($P > 0.03$, Supplementary Table 4).

The other variant associating with cell density is a common allele of a microsatellite at *TCF4* (MAF = 6.1%), a CTG repeat of length ≥ 33, corresponding to the expanded CTG 18.1 allele (OMIM: 602272, allelic variant 0.0007). This microsatellite is pathogenic according to Clinvar and has been reported to strongly predispose to autosomal dominant FECD[35–38], a disease of the corneal endothelium that affects roughly 4% of people over 40 years old (OMIM: 602272). Consistent with the characteristics of FECD, the expanded CTG 18.1 allele associates with lower cell density and HEX ($\beta = -0.38$ SD, $P = 1.6 \times 10^{-19}$ and $\beta = -0.37$ SD, $P = 5.9 \times 10^{-18}$, respectively). Interestingly, it also associates with decreased CH, CRF and IOPg ($\beta = -0.29$ SD, $P = 3.1 \times 10^{-12}$, $\beta = -0.30$ SD, $P = 7.9 \times 10^{-13}$ and $\beta = -0.18$ SD, $P = 1.7 \times 10^{-5}$, respectively) while not affecting glaucoma risk (OR = 0.92, CI = (0.82;1.03), $P = 0.15$). Consistent with previous reports, the expanded CTG 18.1 allele associates with hereditary corneal dystrophies in our data (OR = 7.7, $P = 3.3 \times 10^{-31}$, Supplementary Table 4).

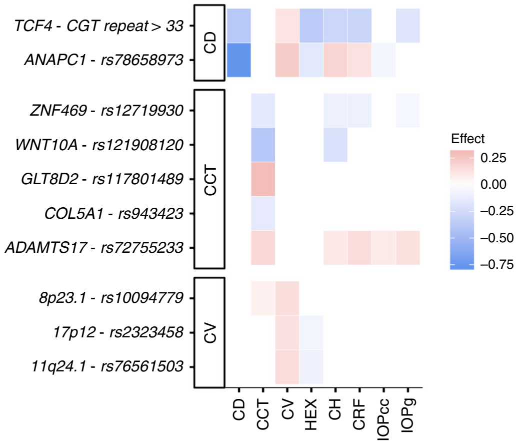

Fig. 3 Effects of the GWS corneal structure variants on different corneal measures. Each row shows the estimated effect of the minor allele on the corneal measures from the specular microscopy and the ocular response analyzer. The variants are annotated with their corresponding gene and grouped by their strongest associating trait. The effect is shown only for significant associations after adjusting for multiple testing with a false discovery rate procedure for each variant. Red color represents a positive effect on the corneal measures and blue color represents a negative effect. Non-significant associations are colored white

The GWAS on HEX revealed only the two variants at *ANAPC1* and *TCF4* (Table 2, Fig. 2b), both of which associate more strongly with cell density.

We identified three loci associating most strongly with CV (Fig. 2c, Table 2). At 17p12, an intergenic variant rs2323458[A] located ~300 kb downstream of *HS3ST3B1*, associates with increased CV (AF = 36.1%, $\beta = 0.14$ SD, $P = 6.9 \times 10^{-13}$). rs2323458 is in moderate linkage disequilibrium (LD) with rs2323457 ($r^2 = 0.61$) which has been reported to associate with CCT[11]. In our data, rs2323457 also associates with CV ($\beta = 0.14$ SD, $P = 1.4 \times 10^{-10}$) but conditional analysis revealed that the effect is driven by rs2323458 ($\beta_{\text{adjusted}} = 0.043$ SD, $P_{\text{adjusted}} = 0.21$). We did not replicate the effect of rs2323457 on CCT ($\beta = -0.033$ SD, $P = 0.13$), even though the power for replication is 94% at a two-sided significance level of 0.05. Two more variants associate with CV: rs10094779[G] at 8p23.1 ~40 kb upstream of *MIR124-1* and rs76561503[C] at 11q24.1 ~60 kb downstream of *MIR100HG* (AF = 24.2%, $\beta = 0.15$ SD, $P = 7.6 \times 10^{-12}$ and AF = 17.9%, $\beta = 0.16$ SD, $P = 3.3 \times 10^{-10}$, respectively). Variants at 8p23.1, between *MIR124-1* and *MSRA*, have been reported to associate with high myopia[39]. rs10094779 is only moderately correlated with the reported variants ($r^2 < 0.25$). Interestingly, a variant in *MIR100HG*, rs577948, has also been reported to associate with myopia[40]. However, rs76561503 is not correlated with the reported variant ($r^2 = 0.03$).

Five variants associated with CCT in our data (Fig. 2d, Table 2). Three are at established CCT loci: *ZNF469*, *WNT10A*, and *COL5A1*[8,9,41]. The two novel CCT associations are with the missense variants p.Thr446Ile in *ADAMTS17* (MAF = 13.8%, $\beta = 0.18$ SD, $P = 1.3 \times 10^{-10}$) and p.Tyr24Cys in *GLT8D2* (MAF = 4.3%, $\beta = 0.30$ SD, $P = 3.9 \times 10^{-10}$). P. Thr446Ile in *ADAMTS17* has been associated with decreased intraocular pressure[42] and decreased height[43]. P.Thr446Ile associates with decreased intraocular pressure in our data ($\beta = 0.14$ SD, $P = 5.8 \times 10^{-7}$), but after adjusting for CCT the association is much weaker ($\beta_{\text{adjusted}} = 0.06$ SD, $P_{\text{adjusted}} = 0.024$). Rare sequence variants in *ADAMTS17* cause autosomal recessive Weill-Marchesani syndrome, a rare connective tissue disorder with features including microspherophakia, severe myopia, glaucoma, cataract, and short stature (OMIM: 607511). Notably, p.Tyr24Cys in *GLT8D2* associates with increased height in a meta-analysis of the Icelandic and UK Biobank data ($\beta = 0.06$ SD, $P = 1.4 \times 10^{-11}$, $N = 490{,}381$). To understand the relationship between height and CCT, we evaluated the correlation between the effect on height in the Icelandic and UK Biobank data ($N = 490{,}381$) of 693 reported adult height variants[44] and their effects on CCT, but found no correlation ($r^2 = 0.006$; $P = 0.038$; F test) (Supplementary Figure 13). To validate the CCT association of the two novel variants, we tested them in a non-overlapping sample of 1459 Icelanders with CCT measurements from the Reykjavik Eye Study[45]. At a significance threshold of $P < 0.05$, we replicated the associations for both p.Thr446Ile in *ADAMTS17* and p.Tyr24Cys in *GLT8D2* with CCT ($\beta = 0.25$ SD, $P = 7.3 \times 10^{-3}$ and $\beta = 0.11$ SD, $P = 0.045$, respectively) (Supplementary Table 5).

Gene expression. We examined the expression levels of the genes at the 10 associating loci using the publicly available Ocular Tissue Database[46]. For non-coding variants we looked for all genes in a 500 kb region around the associating variant. We found that 13 of the 15 genes at the 10 loci are expressed in all ocular tissues (Supplementary Data 1). *MIR124-1* and *MIR100HG* were not found in the database, but previous studies have reported that *MIR124-1* is expressed in the human lens[47] and both *MIR124-1* and *MIR100HG* are expressed in the human retina[40,48]. Expression levels varied across tissues for some genes, e.g., *GLT8D2* is most highly expressed in the cornea, *TCF4* in the sclera and *ANAPC1* in the optic nerve head.

Disease variants affect corneal structure. Two different mutations in *CHST6* are known to cause MCD in Iceland; a missense variant p.Ala128Val (MAF = 0.66%) and a frameshift variant p. Val6MetfsTer106 (MAF = 0.07%)[15]. The prevalence of MCD in Iceland is ~ 1/13,000[49]. We observed that 11 out of 16 homozygous carries of p.Ala128Val, and two out of three homozygous carriers of p.Val6MetfsTer106, have been diagnosed with hereditary corneal dystrophies (ICD10 code H18.5). We investigated the effect of these known disease variants on the corneal measures from the specular microscopy. P.Val6MetfsTer106 associates with cell density and HEX ($\beta = -3.02$ SD, $P = 4.5 \times 10^{-4}$ and $\beta = -2.00$ SD, $P = 9.3 \times 10^{-3}$, respectively) under the recessive model. The association is due to two homozygote carriers showing extremely low values (Supplementary Table 6.a). No homozygote carriers of p.Ala128val participated in the deCODE health study. We did not observe an effect of the two variants among heterozygous carriers (Supplementary Table 6b).

Out of the 6125 study participants, 194 (3.2%) had primary open-angle glaucoma (POAG). Among our corneal measures from the specular microscopy, POAG correlates most strongly with cell density (glaucoma patients have 104 cell/mm^2 lower cell density than controls, $P = 2.1 \times 10^{-6}$; F test). It also correlates with CH, IOPcc and CCT (Table 1).

Due to the correlation between various corneal measures and glaucoma status we assessed the effects of all 15 variants reported to associate with POAG[50] (Supplementary Data 2). First, we replicated the association of 11 out of the 15 variants with POAG ($P < 0.05$). The estimated replication power for the remaining four variants at *GMD2, ZFPM2, ATXN2,* and *PMM2* was > 99.7% at two-sided significance level of 0.05. We investigated the effects of the 11 replicated variants on different corneal measures (Fig. 4a). Five of the 11 variants associate with IOPg, where the POAG risk increasing allele associates with increased IOPg. Even though the strongest relationship of glaucoma is with CH (CH is lower in glaucoma patients), only two of the POAG variants associate with CH. Counter-intuitively, the *TMCO1* allele that increases glaucoma risk associates with lower CH but the *FNDC3B* allele that increases glaucoma risk associates with greater CH. The variant in *FOXC1* is the only POAG variant that also affects cell density. The correlation between the variants' effect on POAG and their effect on available corneal measures were not significant for any trait, controlling the false discovery rate at 0.05 (Supplementary Table 7).

Previously reported variants for CCT and IOP. GWASs have been published for two corneal traits, CCT and IOP. For CCT, 49 variants have been reported[7,11] and we replicate 28 of them ($P <$ 0.05) (Supplementary Data 3). The CCT effects of these variants, and the two novel CCT variants in *ADAMTS17* and *GLTD82*, correlate with their effects on CH, CRF, and IOPg (Supplementary Table 7, Supplementary Figure 14). The effects on different corneal measures by the CCT increasing allele is shown in Fig. 4b. A recent study using participants from the UK Biobank ($N = 115{,}486$) identified 209 variants at 175 novel loci associating with IOPg[51]. We replicate 56 out of these 209 variants in the much smaller Icelandic data ($P < 0.05$) (Supplementary Data 4). The effects on different corneal measures by the IOPg increasing allele is shown in Fig. 4c. The effects of the IOP variants on IOPg correlate with their effects on CH, CRF, IOPcc, and POAG (Supplementary Table 7, Supplementary Figure 15).

Discussion

Using measurements from a specular microscopy, we discovered seven novel variants associating with measurements of corneal structure, i.e., CCT, cell density, CV, and HEX. We further examined their effect on ocular biomechanics, such as CH, CRF, and IOP, as well as examining their effects on the risk of glaucoma and corneal diseases.

The most significant finding is the association of rs78658973 near *ANAPC1* with cell density. *ANAPC1* encodes the Anaphase Promoting Complex Subunit 1, a cell cycle-regulated E3 ubiquitin ligase that controls progression through mitosis and the G1 phase of the cell cycle[52]. The complex is composed of 15–17 subunits, which are highly conserved from yeast to humans. Mutations in the orthologous gene in the fruitfly, *shattered* (*shtd*), result in defective eye development[53] because of disruption of the G1 cell cycle arrest and progression through mitosis. Interestingly, rs78658973 also strongly increases CH, independently of its effect on cell density. CH measures ability of the corneal tissue to dampen pressure changes; such mechanical properties of most tissues are dominated by an extracellular matrix (ECM), in which the connective tissue fibers provide mechanical strength[54]. Many rare connective tissue disorders are characterized by both skeletal and eye abnormalities, such as Marfan syndrome, Ehlers-Danlos syndrome, dermatosparaxis type and Weill-Marchesani syndrome (OMIM: 154700, 225410, and 607511, respectively). An intron variant in *ANAPC1*, rs17040773[G], in complete LD with rs78658973[A] ($r^2 = 1.00$), has been reported to associate with decreased bone mineral density[55] and *ANAPC1* is expressed in bone (Hs.436527). A possible mechanism explaining the association of rs78658973 with cell density in the corneal endothelium may be the role *ANAPC1* has in controlling proliferation of the developing corneal endothelial cells, and likewise, proliferation of bone cells influencing bone density. However, a direct relationship between rs78658973 and *ANAPC1*, or any other gene at the locus, remains to be shown. The estimated fraction of variance of cell density explained by rs78658973 is 24% which is extremely high for such a complex human trait. In comparison, no other variant listed in the GWAS catalog[56] explains higher fraction of variance of a quantitative phenotype available in the extensive deCODE database (Supplementary Table 8). This finding has clinical importance, since reduction in endothelial cell density along with more variable cell size and shape is usually the first sign of corneal endothelial diseases and cell density is important when assessing risk of corneal endothelial failure prior to intraocular surgery[57]. Corneal graft survival studies found cell density lower than 1700 cells/mm^2, 6 months after graft surgery to be associated with increased risk of graft failure[58]. Thirty-year-old homozygote carriers of the *ANAPC1* variant have on average 2637 endothelial cells per mm^2, which is less than the average 70-year old noncarriers (Fig. 1e). Further studies could assess if carriers of the variant are more likely to suffer endothelial decompensation after intraocular surgery or if their cornea is associated with increased risk of corneal graft failure when used as donors.

We also found association of variants at known disease loci, *TCF4* and *CHST6* (FCED and MCD, respectively), with cell density and HEX, demonstrating how these structural measurements are affected by corneal diseases. The *TCF4* variant also strongly associates with CH and it is interesting to note that the *TCF4* and *ANAPC1* variants control almost the same quantitative corneal traits (Fig. 3) but have very different associations with corneal disease. Furthermore, glaucoma risk is strongly associated with lower cell density, while for example the *ANAPC1* variant that controls a quarter of the cell density variance does not associate directly with POAG or PACG. These findings suggest that the pathogenic processes that cause corneal diseases and

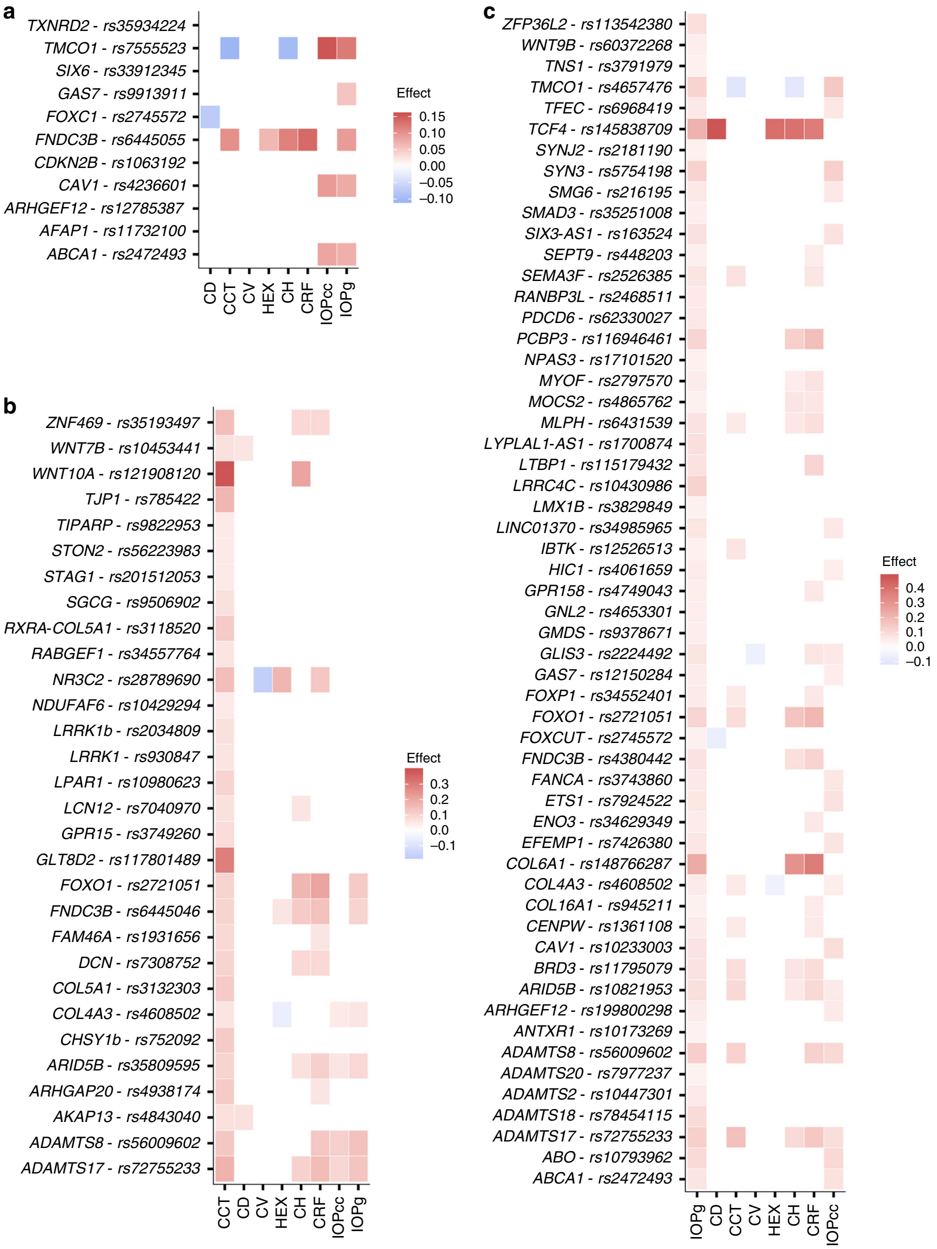

Fig. 4 The effect of reported POAG, CCT, and IOP variants on corneal measures. Red color represents a positive effect on the corneal measures and blue color represents a negative effect. **a** Effect of previously reported POAG variants on corneal traits for the POAG risk increasing allele. Effects on the traits are shown for significant associations after adjusting for multiple testing with a false discovery rate procedure for each variant. **b** Effect of previously reported CCT variants that replicate in our data ($P < 0.05$) and novel CCT variants on corneal measures for the CCT increasing allele. Effects on other traits are shown for significant associations after adjusting for multiple testing with a false discovery rate procedure for each variant. **c** Effect of previously reported IOPg variants that replicate in our data ($P < 0.05$) on corneal measures for the IOPg increasing allele. Effects on other traits are shown for significant associations after adjusting for multiple testing with a false discovery rate procedure for each variant

glaucoma may also lower the cell density, but low cell density in and of itself does not increase risk of disease. In addition, we evaluated the effects of reported POAG variants on corneal structure to further explore the relationship between these traits. Even though CH, cell density, and CCT are lower in POAG patients, the effects of POAG associating variants on POAG risk do not correlate with their effects on these traits in our data. This suggests that these variants do not confer risk of POAG through their effect on these corneal metrics.

Healthy corneal endothelium is necessary for visual perception and its dysfunction is the most common reason for corneal transplantation. Our understanding of the structure and function of corneal endothelial cells and how they relate to diseases is limited. The work presented here constitutes a contribution toward shedding a light on this.

Methods

Study Subjects. Endothelial images from non-contact auto-tracking and -focusing Konan CellCheck SL specular microscopy (Konan Medical USA Inc., Irvine, CA) and measures of ocular biomechanics using the Reichert ocular response analyzer^R (ORA G3, Reichert Technologies, Depew, NY, USA) were obtained for 6266 Icelanders as a part of the deCODE health study. Participation in the deCODE health study also includes an online questionnaire and verbal interviews about health and lifestyle, a number of physical measurements, blood sample collection, and permission to access health-related information from a range of registries and records, including hospital data. We defined subjects in the study to have glaucoma if they reported history of glaucoma or had a hospital discharge diagnosis of primary open-angle glaucoma (ICD10 code H40.1). CCT associations were replicated in a previously described dataset from the Reykjavik Eye Study[59]. Testing variants for association with ocular diseases, we defined the glaucoma and corneal disorder populations based on six different ICD diagnoses; glaucoma information was obtained from participants in the Reykjavik Eye Study and from Icelandic ophthalmologists as described previously[60] (ICD10 H40.1, 4004 cases and 237,214 controls), primary open-angle glaucoma (ICD10 code H40.1, 1261 cases and 303,388 controls), primary angle-closure glaucoma (ICD10 code H40.2, 78 cases and 229,149 controls), disorders of cornea (ICD10 code H18, 133 cases and 255,274 controls), corneal degeneration (ICD10 code H18.4, 81 cases and 287,394 controls), and hereditary corneal dystrophies (ICD code H18.5, 119 cases and 301,665 controls). Written informed consent was obtained from all participants, in accordance with the Declaration of Helsinki, the study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee (VSNb2015120006/03.01 with amendments).

The UK Biobank study is a large prospective cohort study of ~ 500,000 individuals from across UK, aged between 40 and 69 at recruitment[61]. Extensive phenotypic and genotypic information has been collected for the participants, including ICD coded diagnoses from inpatient and outpatient hospital episodes. In this study, we defined the glaucoma and corneal disorder populations based on six different ICD diagnoses; glaucoma (ICD10 code H40, 4428 cases and 404,139 controls), primary open-angle glaucoma (ICD10 code H40.1, 1035 cases and 402,449 controls), primary angle-closure glaucoma (ICD10 code H40.2, 699 cases and 407,868 controls), disorders of cornea (ICD10 code H18, 623 cases and 407,868 controls), corneal degeneration (ICD10 code H18.4, 118 cases and 396,627 controls), and hereditary corneal dystrophies (ICD code H18.5, 211 cases and 378,987 controls). Diagnoses were obtained from primary or secondary diagnoses codes a participant had recorded across all their episodes in hospital. Self-reported diagnoses were excluded from our analysis and we only included individuals determined to be of white British ancestry[62]. We did not exclude related individuals from the analysis but use LD score regression[63] to account for inflation in test statistics due to relatedness. In addition, height measurements were available for 407,825 individuals. UK Biobank's scientific protocol and operational procedures were reviewed and approved by the North West Research Ethics Committee (REC Reference Number: 06/MRE08/65), and informed consent was obtained from all participants.

Images from Konan CellCheck SL specular microscopy. The specular microscopy images were taken by specially trained nurses. One image was taken per eye. The nurse made a visual assessment of the image and selected automated analysis with auto trace for all normal images. Before automatic analysis, the size of cells (S, M, L, or XL) was determined and the cell border lines were compared with the cell borders visually.

All automated analysis were reviewed by a cornea specialist (G.M.Z.) and if found acceptable the automated analysis was used.

Images with any abnormalities, e.g., black areas, highly irregular cell size or shape, or poor quality images, were marked for manual analysis. These images were reviewed and analysed with the Center Method or in a few cases with the Flex Center Method. Images, where no cell structure was seen, were marked ungradeble. All analysis was done by a cornea specialist (G.M.Z.).

We excluded 415 low-quality images for 276 subjects prior to the analysis by regressing cell count against estimated cell density and removed images where the residuals from the fitted model where < -15. Thus, the sample size reduced from $N = 6266$ to $N = 6125$.

Whole-genome sequencing. The process used to whole-genome sequence the 28,075 Icelanders, and the subsequent imputation has been described in a recent publication[30,31]. In summary, we sequenced the whole genomes of 28,075 Icelanders using Illumina technology to a mean depth of at least 10 × (median 32 ×). SNPs and indels were identified and their genotypes called using joint calling with Graphtyper[64]. In total, 155,250 Icelanders were genotyped using Illumina SNP chips and their genotypes were phased using long-range phasing[65]. All sequenced individuals were also chip-typed and long-range phased, which provided information about haplotype sharing that was used to improve genotype calls. Genotypes of the 37.6 million high quality sequence variants were imputed into all chip-typed Icelanders. Using genealogic information, the sequence variants were also imputed into relatives of the chip-typed to further increase the sample size for association analysis and increased the power to detect associations. All of the variants that were tested had imputation information over 0.8.

In UK Biobank genotyping was performed using a custom-made Affimetrix chip, UK BiLEVE Axiom in the first 50,000 individuals[66], and with Affimetrix UK Biobank Axiom array in the remaining participants[67]; 95% of the signals overlap in both chips. Imputation was performed by Wellcome Trust Centre for Human Genetics using a combination of 1000Genomes phase 3[68], UK10K[69] and HRC reference panels[70], for up to 92,693,895 SNPs[62].

Association analysis. All quantitative ocular measurements were averaged for both eyes, rank-based inverse normal transformed to a standard normal distribution separately for each sex and adjusted for age using a generalized additive model[71]. For each sequence variant, a linear regression model, using the genotype as an additive covariate, the transformed quantitative trait as a response and assuming the variance–covariance matrix to be proportional to the kinship matrix, was used to test for association.

We used LD score regression to account for distribution inflation in the dataset due to cryptic relatedness and population stratification[63]. With a set of 1.1 million variants we regressed the χ^2 statistics from our GWASs against LD score and used the intercepts as a correction factors. The estimated correction factors were 1.05, 1.03, 1.03, 1.06, 1.06, 1.05, 1.05, and 1.05 for cell density, CV, HEX, CCT, CH, CRF, IOPg, and IOPcc, respectively.

We used logistic regression to test for association between sequence variants and binary traits, regressing trait status against expected genotype count. In the Icelandic data, we adjusted for sex, age or age at death, county of birth, blood sample availability and an indicator function for the overlap of the subject's lifetime with the time span of phenotype collection, by including these variables in the logistic regression model. In the UK Biobank data, we adjusted for sex and age, as well as 40 principle components in order to adjust for population stratification.

For the meta-analysis we used a fixed-effects inverse variance method[72] based on effect estimates and standard errors from deCODE and the UK Biobank study. Sequence variants from deCODE and the UK Biobank imputation were matched on position and alleles.

Significance thresholds. The thresholds for genome-wide significance were estimated from the Icelandic data and corrected for multiple testing with a weighted Bonferroni adjustment using the enrichment of variant classes as weights with predicted functional impact among association signals[32]. With 37.6 million sequence variants in the Icelandic data, the weights given in Sveinbjornsson et al. were rescaled to control the family-wise error rate. This resulted in significance thresholds of 2.5×10^{-7} for loss-of-function variants, 5.0×10^{-8} for moderate-impact variants, 4.5×10^{-9} for low-impact variants, 2.3×10^{-9} for other variants within DHS sites, and 7.5×10^{-10} for remaining variants. We evaluated false discovery rate, assessed with the q-value package in R. The P value cutoff of 5.0×10^{-8} corresponded to q-values of 0.0014 for cell density, 0.0035 for CV, 0.0117 for CCT, and 0.0116 for HEX, which add up to 3.4%.

When assessing, if associating variants have an effect on other corneal trait, we used the Benjamini–Hochberg false discovery rate (FDR) procedure controlling the FDR at 0.05 at each variant to account for multiple testing.

Correlation between effect sizes. We assessed the relationship between the effects of sequence variant on any two different traits by fitting a weighted linear regression model where the effects sizes for trait 1 was regressed on effect sizes for trait 2 and each variant was weighted by $f(1-f)$ where f is the minor allele frequency of the variants, so that rare variants have less weight in the computation than common variants. For binary traits we used log(OR) as effect size.

Fraction of variance explained. The fraction of variance explained is calculated using the formula $2f(1-f)a^2$ where f is the minor allele frequency and a is the additive effect[73]. Calculating the fraction of variance explained for variants in the GWAS catalog, we estimated the effects of published variants with corresponding phenotypes available in the deCODE data and calculated the fraction of variance explained using *f* and *a* obtained from the Icelandic population.

Code availability

We used the following publicly available software for the whole-genome sequencing process:
for BWA 0.7.10 mem, see https://github.com/lh3/bwa; for Picard tools 1.117, see https://broadinstitute.github.io/picard/; for SAMtools 1.3, see http://samtools.github.io/; for Bedtools v2.25.0-76-g5e7c696z, see https://github.com/arq5x/bedtools2/; for GraphTyper 1.3, see https://github.com/DecodeGenetics/graphtyper; for Variant Effect Predictor, see https://github.com/Ensembl/ensembl-vep.

Data availability

The sequence variants from the Icelandic population whole-genome sequence data have been deposited at the European Variant Archive under accession PRJEB15197, GWAS summary statistics for association with $P < 1 \times 10^{-6}$ are available in Supplementary Data 5. The authors declare that the data supporting the findings of this study are available within the article, it supplementary files, and upon request.

Received: 15 August 2018 Accepted: 28 February 2019
Published online: 20 March 2019

References

1. Ruth, N. & Peralta, V. 2015 Eye Banking Statistical Report. *Eye Bank Assoc. Am.* **3**, 58–69 (2016).
2. Bourne, W. M. Biology of the corneal endothelium in health and disease. *Eye* **17**, 912–918 (2003).
3. Joyce, N. Proliferative capacity of corneal endothelial cells. *Exp. Eye Res.* **95**, 16–23 (2012).
4. Hamuro, J. et al. Cell homogeneity indispensable for regenerative medicine by cultured human corneal endothelial cells. *Investig. Opthalmology Vis. Sci.* **57**, 4749 (2016).
5. Zoega, G. M. et al. Prevalence and risk factors for cornea guttata in the Reykjavik Eye Study. *Ophthalmology* **113**, 565–569 (2006).
6. McCarey, B. E., Edelhauser, H. F. & Lynn, M. J. Review of corneal endothelial specular microscopy for FDA clinical trials of refractive procedures, surgical devices, and new intraocular drugs and solutions. *Cornea* **27**, 1–16 (2008).
7. Gao, X. et al. Genome-wide association study identifies WNT7B as a novel locus for central corneal thickness in Latinos. *Hum. Mol. Genet.* **25**, 5035–5045 (2016).
8. Li, X. et al. Genetic association of COL5A1 variants in keratoconus patients suggests a complex connection between corneal thinning and keratoconus. *Investig. Ophthalmol. Vis. Sci.* **54**, 2696–2704 (2013).
9. Cuellar-Partida, G. et al. WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Hum. Mol. Genet.* **24**, 5060–5068 (2015).
10. Igo, R. P. et al. Differing roles for TCF4 and COL8A2 in Central corneal thickness and Fuchs endothelial corneal dystrophy. *PLoS*ONE **7**, e46742 (2012).
11. Lu, Y. et al. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* **45**, 155–163 (2013).
12. Iglesias, A. I. et al. Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nat. Commun.* **9**, 1864 (2018).
13. Vitart, V. et al. New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. *Hum. Mol. Genet* **19**, 4304–4311 (2010).
14. Afshari, N. A. et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat. Commun.* **8**, 14898 (2017).
15. Liu, N.-P., Smith, C. F., Bowling, B. L., Jonasson, F. & Klintworth, G. K. Macular corneal dystrophy types I and II are caused by distinct mutations in the CHST6 gene in Iceland. *Mol. Vis.* **12**, 1148–52 (2006).
16. Adamis, A. P., Filatov, V. & Tripathi, B. J. Fuchs' endothelial dystrophy of the cornea. *Surv. Ophthalmol.* **38**, 149–168 (1993).
17. Sarnicola, C., Farooq, A. V. & Colby, K. Fuchs endothelial corneal dystrophy: update on pathogenesis and future directions. *Eye Contact Lens* **0**, 1–10 (2018).
18. Ebenesersdóttir, S. S. et al. Ancient genomes from Iceland reveal the making of a human population. *Science* **360**, 1028–1032 (2018).
19. Jonasson, F. et al. Macular corneal dystrophy in Iceland: a clinical, genealogic, and immunohistochemical study of 28 patients. *Ophthalmology* **103**, 1111–1117 (1996).
20. Gagnon, M. M., Boisjoly, H. M., Brunette, I., Charest, M. & Amyot, M. Corneal endothelial cell density in glaucoma. *Cornea* **16**, 314–8 (1997).
21. Gemenetzi, M., Yang, Y. & Lotery, A. J. Current concepts on primary open-angle glaucoma genetics: a contribution to disease pathophysiology and future treatment. *Eye* **26**, 355–369 (2012).
22. Tham, Y. C. et al. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* **121**, 2081–2090 (2014).
23. Cho, S. W., Kim, J. M., Choi, C. Y. & Park, K. H. Changes in corneal endothelial cell density in patients with normal-tension glaucoma. *Jpn. J. Ophthalmol.* **53**, 569–573 (2009).
24. Shah, S., Laiquzzaman, M., Mantry, S. & Cunliffe, I. Ocular response analyser to assess hysteresis and corneal resistance factor in low tension, open angle glaucoma and ocular hypertension. *Clin. Exp. Ophthalmol.* **36**, 508–513 (2008).
25. Shah, S., Laiquzzaman, M., Cunliffe, I. & Mantry, S. The use of the Reichert ocular response analyser to establish the relationship between ocular hysteresis, corneal resistance factor and central corneal thickness in normal eyes. *Contact Lens Anterior Eye* **29**, 257–262 (2006).
26. Zhang, C. et al. Corneal hysteresis and progressive retinal nerve fiber layer loss in glaucoma. *Am. J. Ophthalmol.* **166**, 29–36 (2016).
27. De Moraes, C. V. G., Hill, V., Tello, C., Liebmann, J. M. & Ritch, R. Lower corneal hysteresis is associated with more rapid glaucomatous visual field progression. *J. Glaucoma* **21**, 209–213 (2012).
28. Medeiros, F. A. et al. Corneal hysteresis as a risk factor for glaucoma progression: a prospective longitudinal study. *Ophthalmology* **120**, 1533–1540 (2013).
29. Hoffmann, E. M. et al. Distribution of central corneal thickness and its association with ocular parameters in a large central European cohort: the Gutenberg health study. *PLoS* ONE **8**, e66158 (2013).
30. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
31. Jónsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
32. Sveinbjornsson, G. et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
33. Kotecha, A., Elsheikh, A., Roberts, C. R., Zhu, H. & Garway-Heath, D. F. Corneal thickness- and age-related biomechanical properties of the cornea measured with the ocular response analyzer. *Investig. Ophthalmol. Vis. Sci.* **47**, 5337–5347 (2006).
34. Deol, M., Taylor, D. A. & Radcliffe, N. M. Corneal hysteresis and its relevance to glaucoma. *Curr. Opin. Ophthalmol.* **26**, 96–102 (2015).
35. Mootha, V. V., Gong, X., Ku, H. C. & Xing, C. Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in Fuchs' endothelial corneal dystrophy. *Investig. Ophthalmol. Vis. Sci.* **55**, 33–42 (2014).
36. Wieben, E. D. et al. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy. *PLoS* ONE **7**, e49083 (2012).
37. Riazuddin, S. A. et al. Replication of the TCF4 intronic variant in late-onset Fuchs corneal dystrophy and evidence of independence from the FCD2 locus. *Investig. Ophthalmol. Vis. Sci.* **52**, 2825–2829 (2011).
38. Grant, S. Fa et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).
39. Meng, W. et al. A genome-wide association study provides evidence for association of chromosome 8p23 (MYP10) and 10q21.1 (MYP15) with high myopia in the French population. *Investig. Ophthalmol. Vis. Sci.* **53**, 7983–7988 (2012).
40. Nakanishi, H. et al. A genome-wide association analysis identified a novel susceptible locus for pathological myopia at 11q24.1. *PLoS Genet.* **5**, e1000660 (2009).
41. Yi, L. et al. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* **45**, 155–163 (2013).
42. Choquet, H. et al. A large multi-ethnic genome-wide association study identifies novel genetic loci for intraocular pressure. *Nat. Commun.* **8**, 2108 (2017).
43. Benonisdottir, S. et al. Epigenetic and genetic components of height regulation. *Nat. Commun.* **7**, 13490 (2016).
44. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
45. Jonasson, F. et al. Prevalence of open-angle glaucoma in Iceland: Reykjavik eye study. *Eye* **17**, 747–753 (2003).
46. Wagner, A. H. et al. Exon-level expression profiling of ocular tissues. *Exp. Eye Res.* **111**, 105–111 (2013).

47. Tian, L., Huang, K., Duhadaway, J. B., Prendergast, G. C. & Stambolian, D. Genomic profiling of miRNAs in two human lens cell lines. *Curr. Eye Res.* **35**, 812–818 (2010).
48. Arora, A., McKay, G. J. & Simpson, D. A. C. Prediction and verification of miRNA expression in human and rat retinas. *Investig. Ophthalmol. Vis. Sci.* **48**, 3962–3967 (2007).
49. Jonasson, F., Johannsson, J. H., Garner, A. & Rice, N. S. C. Macular corneal dystrophy in iceland. *Eye* **3**, 446–454 (1989).
50. Shiga, Y. et al. Genetic analysis of Japanese primary open-angle glaucoma patients and clinical characterization of risk alleles near CDKN2B-AS1, SIX6 and GAS7. *PLoS* ONE **12**, e0186678 (2017).
51. Gao, X. R., Huang, H., Nannini, D. R., Fan, F. & Kim, H. Genome-wide association analyses identify new loci influencing intraocular pressure. *Hum. Mol. Genet* **0**, 1–9 (2018).
52. Pines, J. Cubism and the cell cycle: the many faces of the APC/C. *Nat. Rev. Mol. Cell Biol.* **12**, 427–438 (2011).
53. Tanaka-Matakatsu, M., Thomas, B. J. & Du, W. Mutation of the Apc1 homologue shattered disrupts normal eye development by disrupting G1 cell cycle arrest and progression through mitosis. *Dev. Biol.* **309**, 222–235 (2007).
54. Liu, B., McNally, S., Kilpatrick, J. I., Jarvis, S. P. & O'Brien, C. J. Aging and ocular tissue stiffness in glaucoma. *Surv. Ophthalmol.* **63**, 56–74 (2018).
55. Estrada, K., Styrkarsdottir, U. & Evangelou, E. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).
56. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).
57. Gal, R. L. et al. Factors predictive of corneal graft survival in the cornea donor study. *JAMA Ophthalmol.* **133**, 246–254 (2015).
58. Lass, J. H. et al. Endothelial cell density to predict endothelial graft failure after penetrating keratoplasty. *Arch. Ophthalmol.* **128**, 63–69 (2010).
59. Eysteinsson, T. et al. Central corneal thickness, radius of the corneal curvature and intraocular pressure in normal subjects using non-contact techniques: Reykjavik Eye Study. *Acta Ophthalmol.* **80**, 11–15 (2002).
60. Thorleifsson, G. et al. Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. *Nat. Genet.* **42**, 906–909 (2010).
61. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
62. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
63. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
64. Eggertsson, H. P. et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
65. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
66. Wain, L. V. et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).
67. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genom.* **18**, 1–7 (2017).
68. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
69. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
70. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
71. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310 (1986).
72. Mantel, N. & Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**, 719–48 (1959).
73. Gudbjartsson, D. F. et al. Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).

Acknowledgements

We would like to thank the participants of the Icelandic deCODE health study for their valuable contribution to research. We also thank the staff at deCODE genetics recruitment center and core facilities, and all our colleagues, for their important collaboration on this work.

Author contributions

E.V.I., S.B., G.Thorleifsson, P.S., A.O., U.S., G.M.Z., U.T., D.F.G., F.J., H.H. and K.S. designed the study and interpreted the results. E.V.I., S.B., S.K., G.Thorleifsson, G.A.A. and D.F.G. analyzed the data. E.V.I., H.H., G.Thorgeirsson, I.J., G.Thorleifsson, G.A.A., G.M.Z. and F.J. performed recruitment and phenotyping. The manuscript was drafted by E.V.I., S.B., P.S., D.F.G., F.J., H.H. and K.S. All authors contributed to the final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09304-9.

Competing interests: The authors E.V.I., S.B., G.Thorleifsson, P.S, A.O., U.S., S.K., G.A. A. G.Thorgeirsson, I.J., U.T., D.F.G., H.H., and K.S., declare competing financial interests as employees of deCODE genetics/Amgen, Inc. G.M.Z. and F.J. declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks Chiea Chuen Khor and the other anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper III

The genetic architecture of age-related hearing impairment

Authors: Erna V. Ivarsdottir[1,2], Hilma Holm[1], Stefania Benonisdottir[1], Thorhildur Olafsdottir[1], Gardar Sveinbjornsson[1], Gudmar Thorleifsson[1], Hannes P. Eggertsson[1], Gisli H. Halldorsson[1,2], Kristjan E. Hjorleifsson[1,3], Pall Melsted[1,2], Arnaldur Gylfason[1], Gudny A. Arnadottir[1], Asmundur Oddsson[1], Brynjar O. Jensson[1], Aslaug Jonasdottir[1], Adalbjorg Jonasdottir[1], Thorhildur Juliusdottir[1], Lilja Stefansdottir[1], Vinicius Tragante[1], Bjarni V. Halldorsson[1,4], Hannes Petersen[5,6], Gudmundur Thorgeirsson[1,5,7], Unnur Thorsteinsdottir[1,5], Patrick Sulem[1], Ingibjorg Hinriksdottir[8], Ingileif Jonsdottir[1,5,9], Daniel F. Gudbjartsson[1,2,*] and Kari Stefansson[1,5,*].

Affiliations:

[1]deCODE genetics/Amgen, Reykjavik, Iceland.

[2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland.

[3]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA

[4]School of Technology, Reykjavik University, Reykjavik, Iceland

[5]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

[6]Akureyri Hospital, Akureyri, Iceland

[7]Division of Cardiology, Department of Internal Medicine, Landspitali University Hospital, Reykjavik, Iceland

[8]National Institute of Hearing and Speech in Iceland, Reykjavik, Iceland

[9]Department of Immunology, Landspitali University Hospital, Reykjavik, Iceland

*Correspondence to: Daniel F. Gudbjartsson (daniel.gudbjartsson@decode.is) or Kari Stefansson (kari.stefansson@decode.is).

Abstract

Age-related hearing impairment (ARHI) is the most common sensory disorder in older adults. We conducted a genome-wide association meta-analysis of 121,934 ARHI cases and 591,699 controls from Iceland and the UK. We detected associations with 51 sequence variants, under either additive or recessive models. Twenty-one of the variants are novel, of which 13 are rare. Of special interest are a missense variant in *LOXHD1* (MAF=1.96%) and a tandem duplication in *FBF1* covering 4 exons (MAF=0.22%) associating with ARHI (OR=3.7 for homozygotes, P=1.7×10^{-22} and OR=4.2 for heterozygotes, P=5.7×10^{-27}, respectively). We constructed an ARHI genetic risk score (GRS) using common variants. Individuals in the top GRS decile are at 2.5-fold greater risk than those in the bottom decile and develop ARHI 10 years earlier. Furthermore, we found that 13 ARHI variants also associate with tinnitus, and the effects of ARHI variants on ARHI and tinnitus are highly correlated, suggesting that these phenotypes have shared genetic causes. This study sheds a new light on the genetic architecture of ARHI, identifying several novel rare variants in both novel hearing genes and in Mendelian deafness genes, and showing that a common variant GRS can identify individuals with risk comparable to carriers of rare high penetrance variants.

Introduction

Hearing impairment is a common sensory defect, affecting 1-2 out of every 1000 infants and over 50% of people over 80 years old[1,2]. Around 80% of prelingual hearing loss is caused by variants in the sequence of the genome[3], most commonly in the *GJB2* gene encoding the connexin 26 protein involved in inner ear homeostasis[4]. Over 100 genes have been identified that cause prelingual or childhood-onset non-syndromic hearing loss, and 75% of those are inherited in a recessive manner (Hereditary Hearing Loss homepage, URLs).

Less is known about the genetics of age-related hearing impairment (ARHI), defined as a gradual decline of auditory function. ARHI is one of the most common chronic conditions affecting the elderly[5] and is associated with communication difficulties and reduced quality of life[6]. ARHI is usually caused by degeneration of the hair cells in the cochlea. The hair cells are specialized receptors that detect auditory stimuli and convert them into nerve signals that are transmitted to the brain[7]. ARHI can be treated, for instance with hearing aids or cochlear implants in severe cases. The heritability of ARHI has been estimated to be around 50% in twin studies[8]. The genetics of ARHI are complicated by the variability in onset, severity and progression, as well as the effect of environmental factors such as noise-exposure that can lead to hearing impairment[9]. Genome wide association studies (GWAS) on ARHI have been performed[10–16] and a recent study based on UK Biobank self-reported hearing difficulty, reported 44 ARHI loci[17].

The standard type of hearing test is performed with an audiometer that delivers pure tones at different frequencies (measured in hertz (Hz)) and different intensities (measured in decibels hearing level (dB HL)). During the test, a sound is played at frequencies of 0.5, 1, 2, 4, 6 and 8 kHz, and each frequency at different intensity levels. The lowest intensity of sound detection for each individual is defined as their hearing threshold. According to the WHO classification of hearing loss, subjects with a hearing threshold above 25 dB HL are considered to have hearing impairment and the severity of the impairment increases with higher thresholds[18] (Supplementary Table 1). Hearing thresholds at frequencies 0.5, 1, 2 and 4 kHz were used in a pure tone average (PTA). These frequencies represent the range of speech.

Individuals with ARHI are at increased risk of tinnitus, the perception of a sound in the absence of an external sound. These phantom sounds are often described as ringing, buzzing or hissing[19]. Most people experience tinnitus at some point in their life, but for 5-15% of the general population the tinnitus is incessant[20]. Treatment for tinnitus is lacking, even though 1-

3% of individuals experience severe tinnitus affecting their life substantially, including difficulty with concentration and sleep[21]. A twin study estimated the heritability of tinnitus to be 56%[22], yet several genetic studies have failed to find associations of sequence variants with tinnitus[23].

To search for sequence variants associating with ARHI, we performed a GWAS meta-analysis of 121,934 cases and 591,699 controls from two non-overlapping Icelandic datasets and the UK Biobank (UKB). Subsequently, we assessed the effect of ARHI associating variants on tinnitus. Fifty-one independent sequence variants at 45 loci associate with ARHI, 41 under an additive model and 10 under a recessive model. Twenty-one of the associations are novel. Using the association results, we furthermore constructed a GRS for ARHI.

Results

Summary of the data and demographics of ARHI in Iceland

We conducted a GWAS of ARHI in three datasets obtained from the deCODE health study (DHS)[24], the National Institute of Hearing and Speech in Iceland (NIHSI) and the UKB (Figure 1, Supplementary Table 2).

The DHS dataset is based on audiometric measures for 11,484 Icelanders, including 4,140 ARHI cases (PTA>25 dB HL) and 7,344 controls, who are a part of a comprehensive phenotyping of a general population sample enriched for carriers of rare and potentially high impact mutations[25]. The subjects were between 18 and 97 years of age at time of recruitment (43.6% men; mean age = 55.4, SD = 14.5, Supplementary Figure 1.a). The NIHSI is a clinic where patients are referred to for hearing and speech difficulties, and the NIHSI dataset consists of 36,905 audiometric measures of 22,212 Icelanders (55.5% men; mean age = 48.0, SD = 32.4, Supplementary Figure 1.b), of which 43.7% were performed on children (<18 years old). The NIHSI dataset is highly skewed towards those with ARHI, with a prevalence among adults of 73.6% for mild (PTA>25 dB HL), 47.1% for moderate (PTA>40 dB HL), 13.2% for severe (PTA>60 dB HL) and 3.1% for profound (PTA>80 dB HL) hearing impairment. Due to this bias, we defined the 9,619 subjects with PTA above 25 dB HL as ARHI cases and designated 298,609 Icelanders with no available hearing data as population controls (excluding individuals in the DHS dataset). The UKB dataset consists of 108,175 cases with self-reported hearing difficulty and 285,746 controls of white British ancestry, at ages ranging between 40 and 69 years (45.6% men; mean age = 56.5, standard deviation (SD) = 8.1).

The DHS dataset provides an opportunity to analyze the demographics of ARHI in Iceland, although we note that some individuals were recruited based on mutations causing or suspected to cause hearing impairment (Supplementary Table 3). The prevalence of hearing impairment was 36.1% for mild, 7.7% for moderate, 1.1% for severe and 0.1% for profound impairment. In line with previous studies[26,27], the prevalence of moderate hearing impairment at 75 years is 34% for men and 22% for women. The audiometric measures show that hearing declines with age at all frequencies but more drastically at the higher frequencies of 4-8 kHz (Table 1, Supplementary Figure 2). The prevalence of mild hearing impairment reaches 5% shortly after 35 years and increases rapidly with age after 40; 18% at 50 years and 40% at 60 years (Supplementary Figure 3). Consistent with previous reports[28], women are at greater risk of ARHI at low-frequencies (0.5 and 1 kHz), while men are at more risk in the higher frequencies (≥2 kHz) (Table 1). Previous studies have observed an association between ARHI and short stature[29–31]. It has been postulated that the association is due to low levels of insulin-like growth factor-1 (IGF-1), which has a role in the development of the cochlea[29–31]. Performing a logistic regression on mild ARHI (>25 dB HL) against sex, age and height, we observe that reduced height associates with increased risk of ARHI at the lower frequencies

0.5, 1 and 2 kHz (Table 1). After adjusting for height, the association with increased risk of ARHI in women at low-frequencies (0.5 and 1 kHz) is no longer significant (Table 1). This indicates that the greater risk of ARHI at low frequencies for women is driven by the association with reduced height. These results replicate in the NIHSI dataset (Supplementary Table 4).

GWAS meta-analysis

To search for sequence variants associating with ARHI, we performed a meta-analysis of the three GWASs from DHS, NIHSI and UKB, analyzing in total 46.9 million sequence variants under both additive and recessive models (Figure 1). The UKB GWAS was performed on two imputed genotype datasets, one based on variants from the Haplotype Reference Consortium (HRC) reference panel and the other based on variants identified through whole exome sequencing (WES) of 50K study participants (Methods). In total, 55 independent variants at 48 loci satisfied our genome-wide significance thresholds that are dependent on sequence variant annotation[32] (Methods, Supplementary Tables 5, 6 and 7, Supplementary Figure 4). Because we do not restrict the definition of ARHI cases with respect to age at measure or severity, we might detect rare variants in the meta-analysis that are causing prelingual or childhood-onset hearing loss instead of ARHI. Due to this, we used the audiometric measures in the Icelandic datasets to estimate the predicted hearing threshold of the carriers in childhood and observed that 4 of the 55 variants cause prelingual or childhood-onset hearing loss rather than ARHI (Methods, Supplementary Table 7 and Supplementary Note 1). The UKB dataset has the largest sample size of the three datasets and the ARHI associations for the common variants are largely driven by the results from that dataset. Fourteen of the ARHI variants did not associate with ARHI in the Icelandic datasets (P>0.05, Supplementary Table 8). However, the effects show a consistent direction in the three datasets and the pair-wise correlation coefficients between effect sizes are >0.56 (Supplementary Figure 5). Thirty of the associations correspond to previously reported ARHI variants[14,17,33,34]. At 6 of those loci, the previously reported variants are non-coding[17], while we identified missense or splice region variants at these loci (r^2>0.85 between our top variant and the reported variant) (Supplementary Table 6). Thirteen of the novel associations are represented by rare variants, of which five are at novel hearing loci, six are located in Mendelian deafness genes and two are secondary associations at previously reported ARHI loci (Figure 1). Through a gene-based burden test, where rare loss-of-function (LOF) variants (MAF<2%) in the same gene were aggregated and tested together, we identified one additional ARHI gene; *AP1M2*.

Rare variants associating with risk of ARHI

We found 16 ARHI variants that have rare genotypes with large effects (expected genotype frequency (EGF) <1.0%), either in the heterozygous state (N=10, MAF<0.5%) or homozygous state (N=6, MAF<10.0%). Thirteen of those are novel, of which 12 are coding variants (Table 2, Figure 1). In Iceland, 4.9% of the population carries at least one of the 16 rare ARHI genotypes with large effects, and of those carriers that are older than 55, 72% have ARHI compared to 55% of non-carriers (DHS dataset). Overall, carriers of rare ARHI variants have a 2.2-fold ($P=1.0\times10^{-12}$) greater risk of mild hearing impairment than the rest of the population, 3.0-fold ($P=1.4\times10^{-9}$) greater risk of moderate and 5.6-fold ($P=1.9\times10^{-8}$) greater risk of severe impairment (DHS dataset, Figure 2.a).

Five of the novel variants that have rare genotypes with large effects are at loci that have not been reported for any type of hearing impairment: *FBF1*, *FSCN2*, *C10orf90*, *SH2D4B* and *TBX2* (Supplementary Note 2).

A rare tandem duplication in *FBF1*, only detected in Iceland, associates strongly with ARHI (MAF_{Ice} = 0.22%, OR=4.2, $P=5.7\times10^{-27}$). The variant is highly penetrant, with 81.5% of the 162 carriers having at least mild ARHI and 57.4% having moderate to profound hearing impairment (Figure 3.a). The duplication spans 7,282 base pairs covering exons 4 to 7 of *FBF1*. To investigate the effects of the duplication on the transcription of the gene, we analyzed RNA sequencing data from whole-blood of heterozygous carriers and non-carriers (N=13,067) (Supplementary Note 3). Out of 60 heterozygous carriers we found evidence of transcripts containing duplication of exons 4 to 7 demarked by an extra splicing between exon 7 and 4 (Figure 3.b). This transcript isoform was not detected in RNA sequences from any of the 13,007 non-carriers. We estimate that transcripts with splicing between exon 4 to 7 represents 7.5% (95% CI 5.7-9.2) of *FBF1* transcripts in carriers. Three other variants are correlated (r^2>0.8) with the duplication but none of them are coding (Supplementary Figure 6.a).

Six of the novel variants with rare genotypes and large effects on ARHI, are coding variants located in Mendelian deafness genes: *LOXHD1*, *MPZL2*, *SLC411*, *SLC26A5*, *TBC1D24* and *TMPRSS3*. Rare variants in these genes have been reported to cause severe to profound hearing impairment described as either prelingual or childhood-onset (DFNB77, DFNB111, DFNB61, DFNA65, DFNB86 and DFNB8; OMIM #613079, #618145, #217400, #613865, #616044, #614617 and #601072). However, apart from *TMPRSS3*, the ARHI associations we find in these genes are with missense variants that have milder effect than the prelingual or childhood-onset variants (Supplementary Note 4, Table 2, Figure 4.a-d).

The missense variant in *LOXHD1*, p.Arg1090Gln, on chromosome 18q21.1, associates with increased risk of ARHI under the recessive model (OR=3.92, P=8.9×10^{-22}). P.Arg1090Gln has a MAF of 2.96% in Iceland and 1.99% in the UK and is the only novel variant that reaches genome-wide significance in both Iceland and the UK (Supplementary Table 6). No other variants are correlated with p.Arg1090Gln (r^2<0.4, Supplementary Figure 6.b). The variant has high penetrance, with 82.3% of the 62 homozygotes in the Icelandic datasets having ARHI and 48.4% having moderate to profound hearing impairment (Figure 4.a). We had information on hearing aid usage and the age when hearing aid usage started for 8,211 of the 11,484 individuals in the DHS dataset and found that 15 out of 37 homozygous carriers use hearing aids and started using them earlier (mean age=46.3 years, SD=15.7) than heterozygotes and non-carriers (mean age=60.0 years, SD=14.8, P=0.031). Heterozygotes are also at increased risk of ARHI (OR=1.07, P=2.0×10^{-4}). However, this risk is much lower than the risk of homozygotes and therefore the effect of this variant deviates from the additive model (P=3.3×10^{-12}, Supplementary Table 9).

A rare frameshift variant, c.208delC, in *TMPRSS3* associates with ARHI under the additive model (OR=1.49, P=8.3×10^{-8}, MAF$_{\text{Ice}}$=0.22%, MAF$_{\text{UK}}$=0.07%). In the homozygous state, c.208delC has been reported to cause congenital deafness (OMIM # 601072), but the increased risk of ARHI of heterozygous carriers has not been reported. In the Icelandic datasets, the only homozygous carrier has profound hearing loss. In the heterozygous state, the variant shows variable expressivity (Figure 4.e); where some carriers have normal hearing, 52 carriers have moderate, 17 have severe and 4 have profound hearing loss. Battelino et al. reported a Slovenian family-trio with congenital profound hearing loss, where the mother and the son were homozygous carriers of c.208delC and the father was a heterozygous carrier of c.208delC in *TMPRSS3* and c.35delG in *GJB2*[35]. Our results suggest that one copy of c.208delC in *TMPRSS3* can cause profound hearing loss.

A novel ARHI gene detected with a burden test

Using a LOF variant gene-based burden test, the gene *AP1M2*, on chromosome 19p13.2, associates with ARHI under the recessive model in the Icelandic datasets (OR=28.9, P=4.6×10^{-7}). Twelve homozygotes or compound-heterozygotes for LOF variants with MAF<2% in *AP1M2* had been invited to participate in the deCODE health study. Nine of them participated, three homozygous carriers of a stop gained variant, p.Arg386Ter

(MAF=0.22%), three homozygous carriers for a splice donor variant, c.673+2T>C (MAF=0.37%), and three compound-heterozygous carriers of these two variants. One additional compound-heterozygous carrier had audiometric measures from the NIHSI. Four of the individuals have severe, two have moderate and two have mild ARHI. Five of them reported use of hearing aids, and their average age when starting using hearing aids (mean=27.8, SD=14.9) is substantially younger than that of other hearing aid users (mean=60.2, SD=15.1, P=0.022). RNA and protein expression analyses of inner-ear tissue[36–38] have shown that Ap1m2 is expressed 7-fold higher in hair cells than in non-hair cells in mice with a FDR of 3.4×10^{-3}, suggesting a specific role in hair cell function (Supplementary Table 10). Non-syndromic hearing loss has been linked to chromosome 19p13.2 in families from Pakistan[39] and Germany[40], without a specific gene being implicated.

The effect of the variants on ARHI by genotype

It is interesting that 20% of the variants associate with ARHI under the recessive model, a much higher fraction than other age-related diseases. To further explore the effect of all of the ARHI associating variants per genotype we tested them under the genotypic model estimating the effect of heterozygous and homozygous carriers separately (Supplementary Table 9). We found that p.Arg402Gln in *TYR*, p.Val504Met in *KLHDC7B*, p.Thr656Met in *SYNJ2* and p.Leu113Val in *CLRN2* have stronger effects on homozygotes than expected under the additive model (P<0.05). Furthermore, we found that the variants in *ILDR1*, *CHMP4C* and *CCDC68*, reported before as additive[17], are better explained by the recessive model, only showing significant effects on homozygous carriers (Supplementary Table 9).

Dimensions of the audiometric data

In the Icelandic datasets, the subjects underwent an audiometric test providing more information about the severity of the ARHI and the affected frequencies than in the UKB dataset. To further explore which hearing frequencies are affected by the ARHI variants, we tested each frequency (0.5, 1, 2, 4, 6 and 8 kHz) separately for association with the ARHI variants (Supplementary Table 8, Figure 5). Most of the ARHI variants have similar effects at all frequencies although some variants have stronger effects on lower frequencies and others on higher frequencies. For instance, p.Arg1090Gln in *LOXHD1* affects the lower frequencies more than higher frequencies under a recessive model, with the greatest effect on 1 kHz (OR=8.4, P=1.6×10^{-18}), which is different from its effect on ARHI at 6 and 8 kHz (P_{het} < 0.02). Furthermore, six ARHI variants, that do not associate with PTA based ARHI in Iceland (P>0.05), associate nominally with ARHI for some particular frequency (Supplementary Table 8, Figure 5).

Association of ARHI variants with tinnitus

We tested the ARHI variants for association with tinnitus using self-reported information from DHS and UKB (N_{cases}=47,657, $N_{controls}$=111,607, Supplementary Table 11). ARHI variants detected under the additive model were tested for tinnitus using the additive model and ARHI variants detected under the recessive model were tested for tinnitus using the recessive model. Thirteen ARHI variants associate with tinnitus, controlling the false discovery rate at 0.05 using the Benjamini-Hochberg procedure (Figure 6, Supplementary Table 11). Variants in *CTBP2, CRIP3*, *AGO2, PHLDB1*, *LMX1A*, *SLC26A5*, *ACADVL*, *SYNJ2* and *CLRN2* associated with tinnitus under the additive model and variants in *ILDR1*, *ABCC10, SH2D4B* and *C10orf90* associated with tinnitus under the recessive model. For all of the 13 variants, the ARHI risk increasing allele increases the risk of tinnitus, and the effect of all the ARHI variants on ARHI risk and tinnitus risk are highly correlated (r = 0.72, P = 6.2×10^{-8} and r = 0.86, P = 6.0×10^{-4} for the additive and recessive model respectively, Figure 6).

Genetic risk score predicts ARHI risk

We constructed a genetic risk score (GRS) for ARHI, based on the 35 ARHI variants with EGF>1%, using effect sizes from the UKB dataset. The GRS associates with ARHI in both Icelandic datasets (OR=1.31, P=4.1×10^{-29} and OR=1.18, P=7.5×10^{-39} in DHS and NIHSI datasets respectively) and the association is dose-dependent over GRS deciles (Figure 2.b). In the DHS dataset, individuals in the top decile of the GRS have 2.5-fold (P=6.1×10^{-18}) greater risk of ARHI than those in the bottom decile. Comparing the cumulative risk of ARHI against age between the top and bottom GRS deciles, shows that individuals in the bottom decile have their ARHI 10 years later than those in the top decile (Figure 2.c). Furthermore, individuals in the top GRS decile have a 3.2-fold (P=2.1×10^{-8}) and 2.7-fold (P=0.031) greater risk of moderate and severe hearing impairment respectively, than those in the bottom decile. If we compare the 4.9% who carry any of the 16 rare ARHI variants to the bottom 10% of the GRS, the ORs are 3.4 for mild, 6.1 for moderate and 9.2 for severe hearing impairment (P=3.0×10^{-19}, 8.4×10^{-13} and 1.0×10^{-7}, respectively). Therefore, relative to the bottom GRS decile, the ARHI OR for carriers of rare variants is larger than the ARHI OR for individuals in the top GRS decile, but the ORs do not show significant heterogeneity (P_{het}=0.075). However, the risk of moderate and severe ARHI for carriers of rare variants is substantially greater than the risk for the top GRS decile (P_{het}<0.05).

As we have described, the severity of the hearing impairment for carriers of the highly penetrant variants in *LOXHD1* and *FBF1* varies from mild to profound. We hypothesize that the GRS could act as a modifier on the expressivity, i.e. that some of the variable expressivity of these highly penetrant variants could be explained by the common variants associating with ARHI. We estimated the relationship between these variants and the GRS on the PTA hearing thresholds and found positive interaction for both *LOXHD1* and *FBF1* (P=6.8×10^{-4} and P=4.7×10^{-3}, respectively). This shows that among carriers of these highly penetrant genotypes, those who additionally have a high GRS are at a greater risk of a more severe ARHI than those that have a low GRS.

Long-term exposure to occupational loud noises is a risk factor for ARHI[9,41]. We had information on the occupation the subjects had for the majority of their lives for 7,642 of the 11,484 individual in the DHS dataset. Three occupational categories associated with increased risk of ARHI; plant and machine operators and assemblers (N=508, OR=1.88, $P=8.4\times10^{-8}$), craft and related trades workers (N=1,172, OR=1.56, $P=1.3\times10^{-7}$) and agricultural and fishery workers (N=783, OR=1.55, $P=1.6\times10^{-6}$). We tested for an interaction effect between long-term occupational noise exposure and the ARHI GRS on the risk of ARHI but did not find a significant interaction (P=0.94). Among the individuals in the top GRS decile, noise exposure associates with increased risk of ARHI (OR=1.77, $P=6.3\times10^{-3}$) similar to the rest of the population (1.70, $P=1.2\times10^{-12}$, $P_{het}=0.86$).

Discussion

In the largest GWAS meta-analysis on ARHI to date, we found association with 51 variants, of which 21 are novel, using audiometric measurements from Icelanders and data on self-reported hearing difficulty from the UKB. This study yielded more rare variants, both under additive and recessive models, than previous GWAS studies that have reported common variants associations with small to moderate effects on ARHI. The novel findings include variants in both known Mendelian deafness genes and genes not previously linked to hearing.

We constructed an ARHI GRS and found that individuals in the top GRS decile are at 2.5-fold greater risk than those in the bottom decile, and on average, they develop ARHI 10 years earlier than the bottom decile. The 2.5-fold greater risk is comparable to the 3.4-fold greater risk of carriers of rare ARHI variants relative to the risk of those in the bottom GRS decile. However, carriers of rare ARHI variants have substantially greater risk of moderate and severe ARHI than individuals in the top GRS decile, showing that the rare variants identified in this study predispose to more severe ARHI than the combination of common variants in the GRS.

Despite the importance of hearing in everyday life, ARHI is often not recognized by patients and left untreated; only 22% of people with mild hearing impairment report a hearing handicap[6]. Because of this, ARHI is often not diagnosed until several years after onset and has then often already had many negative consequences such as effects on employment, social isolation and depressive symptoms[42]. Due to this, there is a need for better screening strategies, and using a GRS to stratify individuals into risk groups could enable enhanced screening. Identifying high risk individuals might also help with preventing or reducing the severity of the hearing impairment. We have shown that noise exposure increases the risk of ARHI among those that are already at a high genetic risk. It shows that avoiding loud noises is even more important for those who have a genetic predisposition to ARHI.

Previous reports have claimed that over 70% of non-syndromic prelingual hearing loss is inherited in a recessive manner[43]. In this study, we found six novel variants that associate with ARHI under a recessive mode of inheritance. For instance, the variant in *LOXHD1* is genome-wide significant in the UKB data alone, but was not detected by Wells et al. with the same dataset under an additive model[17]. Additionally, we show that three variants previously reported to associate with ARHI under an additive model are truly recessive and four variants detected under an additive model in this study have stronger effects on homozygous carriers than expected under an additive model. These results highlight the importance of applying a recessive model when searching for variants associating with ARHI, which has not been done in previous GWASs.

A limitation of this work is that in the three datasets, the ARHI phenotype is defined in different ways. Because we did not restrict the definition of ARHI in terms of age of onset or severity, we performed follow-up analysis for all rare variants to make sure that the reported variants really associate with ARHI and not prelingual or childhood-onset deafness. Our results for common variants were mainly driven by the UK biobank dataset but 38 out of 51 variants replicated in the combined Icelandic datasets. The lack of replication in the DHS dataset is most likely due to smaller sample size, while in the NIHSI dataset it might be due to differences in the phenotype ascertainment, where patients are referred to NIHSI for hearing problems. Using population controls in the NIHSI dataset that have not been specifically screened for hearing impairment, will also misclassify some cases as controls. However, the effect sizes from the UK are highly correlated with effect sizes from Iceland and have consistent direction of effects. The age of onset, severity and progression of ARHI is highly variable between individuals, and future GWAS could further analyze subtypes of ARHI. The Icelandic datasets provide more details regarding these factors as well as the measures of hearing at specific frequencies. Some ARHI variants have stronger effects on particular frequencies, while most affect all frequencies similarly.

We found six loci that have not been reported to affect hearing in humans before; *FBF1*, *FSCN2*, *TBX2*, *C10orf90*, *SH2D4B* and *AP1M2*. Inner-ear protein expression analysis in mice[36] have shown that the mouse homologs Fscn2, Tbx2, C10orf90, Sh2d4b and Ap1m2 have higher expression, ranging from 5 to 43-fold, in hair cells versus non-hair cells (Supplementary Table 10), suggesting that these genes have specialized roles in the inner ear hair cells, but degeneration of the inner ear hair cells is the main cause of ARHI[5].

The two strongest novel associations were with the highly penetrant tandem duplication covering exons 4 to 7 in *FBF1*, detected under an additive model, and a missense variant in *LOXHD1*, affecting ARHI in homozygous state. *FBF1* encodes Fas-binding factor 1, a keratin-binding protein necessary for ciliogenesis[44,45]. We speculate that *FBF1* may have a role in the cilia of the inner ear, but further studies are needed to determine the biological effect of the duplication and the mechanism behind the association of *FBF1* with ARHI. *LOXHD1* encodes lipoxygenase homology domain 1, which consists of 15 PLAT (polycystin-1, lipoxygenase, alpha-toxin) domains[46]. Grillet et al. showed that *LOXHD1* is expressed in the functionally mature mechanosensory hair cells in the inner ear and LOF mutations in the gene lead to auditory defects in mice and humans, indicating an essential role for normal hair cell function[46]. Homozygous carriers of p.Arg1090Gln in *LOXHD1* report a younger age for hearing aid usage than the rest of hearing aid users, showing that the variant is associated with a severe form of ARHI. Given the frequency of p.Arg1090Gln (gnomAD, URLs), we estimate the number of homozygous carriers to be around 300 in Iceland, 24,000 in the UK and 300,000 in the whole of Europe.

We also tested the ARHI variants for association with tinnitus. Tinnitus is considered to have a broad etiology and can be caused by problems in the entire auditory pathway[47]. ARHI and tinnitus are correlated phenotypes, but shared genetic causes have not been broadly explored. We found that 13 of the 51 ARHI variants also associated with tinnitus, showing that some pathogenic processes that cause ARHI also increase the risk of tinnitus.

Our knowledge of the pathogenesis of ARHI is still limited, but the work presented here reveals several novel loci, shedding a new light on the genetics underlying this common sensory defect.

Methods

Phenotype datasets

For the meta-analysis we conducted a GWAS of ARHI in three datasets under both additive and recessive models.

The DHS dataset. Pure tone audiometric air conduction testing was performed for 11,484 Icelanders as a part of a comprehensive phenotyping of a general population sample enriched for carriers of rare and potentially high impact mutations[25] (the deCODE health study[24]). Homozygous carriers of p.Arg1090Gln in *LOXHD1* were recruited, resulting in 37 carriers that participated (Supplementary Table 3). For the GWAS, 4,140 individual with PTA>25 were defined as ARHI cases and the remaining 7,344 as controls. Participation in the deCODE health study includes blood sample collection, numerous physical measurements, permission to access a wide range of health-related information including hospital data, a verbal interview and an online questionnaire about health and lifestyle, including questions on hearing aid usage and tinnitus. All participants of the study gave written informed consent, in accordance with the Declaration of Helsinki, and the study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee (VSNb2015120006/03.01 with amendments).

The NIHSI dataset. Pure tone audiometric air conduction testing was performed for 22,212 Icelanders at the National Institute of Hearing and Speech in Iceland (NIHSI). For the GWAS, 9,619 individuals were defined as ARHI cases (PTA>25) and 298,609 individuals were selected as population controls (excluding individuals in the DHS dataset). All participants who donated samples gave informed consent and the study was approved by the Icelandic Data Protection Authority and National Bioethics Committee (VSN-18-186).

The UKB dataset. The UKB study is a large prospective cohort study of around 500,000 individuals from the UK[48]. Extensive phenotypic and genotypic information has been collected for the participants, including self-reported hearing difficulty and tinnitus. For the GWAS, we defined 108,175 ARHI cases as those who answered “Yes” or “I am completely deaf” to the question “Do you have any difficulty with your hearing?” and 285,746 controls as those who answered “No”. In our analysis we only included individuals determined to be of white British ancestry[49] and we use LD score regression[50] to account for inflation in test statistics due to relatedness. All participants of the UK Biobank study gave informed consent and the study was approved by the North West Research Ethics Committee (REC Reference Number: 06/MRE08/65).

Audiometric test

In DHS and NIHSI datasets, the pure tone air conduction audiometric test was performed by specially trained staff. The audiometer delivers pure tones at 0.5, 1, 2, 4, 6 and 8 kHz at different intensity levels, usually starting at 20 dB HL and increased if necessary. For each individual and each ear, the lowest intensity of sound detection is defined as their hearing threshold at that frequency. The pure tone average (PTA) was defined as the average hearing threshold at 0.5, 1, 2 and 4 kHz (according to the classification of the WHO). We define ARHI cases as those with PTA>25.

Genotype datasets

In the Icelandic GWASs, using the DHS and NIHSI datasets, we analyzed high quality 34.0 million sequence variants identified through whole-genome sequencing of 49,708 Icelanders which have been described in detail[51,52]. In summary, we whole-genome sequenced the Icelanders using Illumina technology to a mean depth of at least 17.8 × and median depth of 36.9 ×. The sequence variants were jointly called using Graphtyper[53], thereof 79.318 high-confidence structural variants described previously[54]. We genotyped 166,281 Icelanders using Illumina SNP chips and their genotypes were phased using long-range phasing[55]. Genotypes of the 34.0 million sequence variants were imputed into all chip-typed Icelanders as well as relatives of the chip-typed, to increase the sample size for association analysis. All tested variants had imputation information over 0.8.

The UKB GWAS was performed with two sets of genotypes. The primary analysis was performed with 26.5 million high quality variants (imputation info > 0.8) from the Haplotype Reference Consortium (HRC) reference panel, imputed into chip-typed individuals of European ancestry[49]. The genotyping was performed using a custom-made Affimetrix chip, UK BiLEVE Axiom in the first 50,000 individuals[56], and with Affimetrix UK Biobank Axiom array in the remaining participants[57]. Imputation was carried out by Wellcome Trust Centre for Human Genetics using a combination of 1000Genomes phase 3[58], UK10K[59] and HRC reference panels[60], for up to 93 million variants[49]. Additionally, we performed a GWAS with 922 thousand variants identified through whole exome sequencing (WES) of 49,960 study participants[61], imputed into chip-typed individuals of European ancestry.

Association analysis

Logistic regression was used to test for association between sequence variants and binary traits. For the additive model, the expected allele counts were used as a covariate while for the recessive model, the product of the maternal and paternal genotype probabilities were used as a covariate. For the genotypic model, separate parameters were included for heterozygotes and homozygotes. Other available individual characteristics that correlate with the trait were additionally included in the model. In the DHS and NIHSI datasets, those were sex, county of birth, current age or age at death (including first and second order terms), blood sample availability and an indicator function for the overlap of the lifetime of the individual with the time span of phenotype collection. In the UKB dataset, those were sex, age and 40 principal components in order to adjust for population stratification.

We used LD score regression to account for distribution inflation in the dataset due to cryptic relatedness and population stratification[50]. Using 1.1 million variants, we regressed the $\chi2$ statistics from our GWASs against LD score and used the intercepts as a correction factors. The estimated correction factors for ARHI were 1.05, 1.20 and 1.05 in DHS, NIHSI and UKB datasets respectively.

Because the UKB GWAS was performed on two sets of genotypes, we performed two separate meta-analysis. Both meta-analysis combined results from three GWAS using DHS, NIHSI and UKB datasets. In meta-analysis I, we used the UKB GWAS results based on the variants from the HRC reference panel and in meta-analysis II we used the UKB GWAS results based on the variants identified through WES. When meta-analyzing the three GWASs, we used a fixed-effects inverse variance method[62] which is based on effect estimates and standard errors from all datasets. Sequence variants from Iceland and the UKB were matched on position and alleles.

For the genotypic model P-values were computed by comparing the genotypic model to the null model. For the genotypic model meta-analysis, sample size approach was used based on P-values and sample size[63].

A Q-test[64] was used to test for heterogeneity between effect sizes.

Definition of ARHI variants

The PTA based definition of ARHI used in the Icelandic datasets does not exclude individuals that are completely deaf or have child-hood onset hearing loss. The GWAS can therefore detect rare associating variants that cause prelingual or childhood onset hearing loss instead of ARHI. Due to this, for all the rare variants that satisfied the genome-wide significance thresholds, we fit a linear regression model, with the PTA hearing threshold of the carriers as response and age as covariate, to estimate the predicted PTA hearing threshold of the carriers in childhood. Variants that had predicted hearing threshold of 25 dB HL at 10 years of age were considered to be causing childhood-onset hearing loss.

Significance thresholds

The genome-wide significance thresholds were corrected for multiple testing with a weighted Bonferroni adjustment[32]. The weights, based on enrichment of variant classes with predicted functional impact among association signals, were estimated from the Icelandic data, resulting in significance thresholds of 2.4×10^{-7} for loss-of-function variants, 4.9×10^{-8} for moderate-impact variants, 4.4×10^{-9} for low-impact variants, 2.2×10^{-9} for other variants within DHS sites and 7.4×10^{-10} for remaining variants.

We evaluated false discovery rate, assessed with the qvalue package in R. The P value cutoff of 5.0×10^{-8} corresponded to q-values of 0.0013 for the additive model and 0.0025 for the recessive model, which add up to 0.4%.

In the burden test, a genome-wide significance threshold of $0.05/18{,}482 = 2.7\times10^{-6}$ was used, correcting for the number of autosomal protein coding RefSeq genes[65,66].

Conditional analysis

To search for secondary association signals at each locus, we applied a stepwise conditional analysis, adding the top variant as a covariate when testing all other variants in a 1 Mb window around the top variant. We used a Bonferroni adjusted significance threshold for secondary associations. We found independent secondary associations at 6 loci.

RNA-sequencing

The RNA sequencing from whole blood of 13,067 Icelanders has been described in previous publications[67,68].

Genetic risk score

The GRS for ARHI was constructed using the 35 detected variants with EGF>1% and estimated effects from the UKB dataset. If we let m_{vi} and p_{vi} be the genotype probability for individual *i* and sequence variant *v* at the maternally and paternally inherited chromosomes, the GRS for individual *i* is defined as

$$grs_i = \sum_{v=1}^{n} (m_{vi} + p_{vi})\beta_v + \sum_{v=1}^{m} (m_{vi} \times p_{vi})\gamma_v,$$

where β are the effects of the *n* variants detected with the additive model and γ are the effects of the *m* variants detected with the recessive model.

Correlation between effect sizes

When assessing the relationship between effect sizes, we fitted a weighted linear regression model where each variant was weighted by $f(1-f)$ where f is the minor allele frequency of the variants.

Additional information

URLs

https://hereditaryhearingloss.org/

https://gnomad.broadinstitute.org/

Acknowledgements

We would like to thank all study participants for their valuable contribution to research. We also thank the staff at NIHSI and deCODE genetics recruitment center and core facilities, and all our colleagues, for their important collaboration on this work. A part of this research has been conducted using the UK Biobank Recourse under application number 56270.

Data Availability Statement

The sequence variants from the Icelandic population whole-genome sequence data have been deposited at the European Variant Archive under accession PRJEB15197, GWAS summary statistics for association with $P < 1 \times 10^{-6}$ are available in Supplementary Data. The authors declare that the data supporting the findings of this study are available within the article, it supplementary files, and upon request.

Author Contributions

E.V.I., H.H., S.B., U.T., P.S., I.H., I.J., D.F.G. and K.S. designed the study and interpreted the results. E.V.I. S.B., G.S., G.Thorleifsson, H.P.E., G.H.H., K.E.H., P.M., A.G., A.O., G.A.A., B.O.J., L.S., B.H. and D.F.G analyzed the data. As.J. and Ad.J. did the Sanger sequencing. E.V.I., H.H., G.Thorleifsson, T.J., V.T., H.P., G.Thorgeirsson, I.H. and I.J. performed recruitment and phenotyping. The manuscript was drafted by E.V.I., H.H., S.B., T.O., U.T., P.S., I.H., D.F.G. and K.S. All authors contributed to the final version of the manuscript.

Competing Financial Interest

E.V.I., H.H., S.B., T.O., G.S., G.Thorleifsson, H.P.E., G.H.H., K.E.H., P.M., A.G., G.A.A., A.O., B.O.J., As.J., Ad.J., T.J., L.S., V.T., B.V.H., G.Thorgeirsson., U.T., P.S., I.J., D.F.G and K.S. are employees of deCODE genetics/Amgen, Inc. H.P. and I.H. have no financial interest to declare.

References

1. Venkatesh, M. D., Moorchung, N. & Puri, B. Genetics of non syndromic hearing loss. *Medical Journal Armed Forces India* **71**, 363–368 (2015).

2. Vos, B., Noll, D., Pigeon, M., Bagatto, M. & Fitzpatrick, E. M. Risk factors for hearing loss in children: a systematic literature review and meta-analysis protocol. *Syst. Rev.* **8**, (2019).

3. Shearer, A. E., Hildebrand, M. S. & Smith, R. J. *Hereditary Hearing Loss and Deafness Overview*. *GeneReviews®* (1993).

4. Snoeckx, R. L. *et al.* GJB2 Mutations and Degree of Hearing Loss: A Multicenter Study. *Am. J. Hum. Genet.* **77**, 945–957 (2005).

5. Yamasoba, T. *et al.* Current concepts in age-related hearing loss: Epidemiology and mechanistic pathways. *Hearing Research* **303**, 30–38 (2013).

6. Dalton, D. S. *et al.* The Impact of Treated Hearing Loss on Quality of Life. *Gerontologist* **43**, 661–668 (2003).

7. Frolenkov, G. I., Belyantseva, I. A., Friedman, T. B. & Griffith, A. J. Genetic insights into the morphogenesis of inner ear hair cells. *Nature Reviews Genetics* **5**, 489–498 (2004).

8. Karlsson, K. K., Harris, J. R. & Svartengren, M. Description and primary results from an audiometric study of male twins. *Ear Hear.* **18**, 114–120 (1997).

9. Kujawa, S. G. Acceleration of Age-Related Hearing Loss by Early Noise Exposure: Evidence of a Misspent Youth. *J. Neurosci.* **26**, 2115–2123 (2006).

10. Friedman, R. A. *et al.* GRM7 variants confer susceptibility to age-related hearing impairment. *Hum. Mol. Genet.* **18**, 785–796 (2009).

11. Fransen, E. *et al.* Genome-wide association analysis demonstrates the highly polygenic character of age-related hearing impairment. *Eur. J. Hum. Genet.* **23**, 110–115 (2015).

12. Girotto, G. *et al.* Hearing function and thresholds: A genome-wide association study in European isolated populations identifies new loci and pathways. *J. Med. Genet.* **48**, 369–374 (2011).

13. Van Laer, L. *et al.* A genome-wide association study for age-related hearing impairment in the Saami. *Eur. J. Hum. Genet.* **18**, 685–693 (2010).

14. Hoffmann, T. J. *et al.* A Large Genome-Wide Association Study of Age-Related Hearing Impairment Using Electronic Health Records. *PLoS Genet.* **12**, (2016).

15. Vuckovic, D. *et al.* Genome-wide association analysis on normal hearing function identifies PCDH20 and SLC28A3 as candidates for hearing function and loss. *Hum. Mol. Genet.* **24**, 5655–5664 (2015).

16. Nagtegaal, A. P. *et al.* Genome-wide association meta-analysis identifies five novel loci for age-related hearing impairment. *Sci. Rep.* **9**, 1–10 (2019).

17. Wells, H. R. R. *et al.* GWAS Identifies 44 Independent Associated Genomic Loci for Self-Reported Adult Hearing Difficulty in UK Biobank. *Am. J. Hum. Genet.* **105**, 788–802 (2019).

18. Mathers, C., Smith, A. & Concha, M. Global burden of hearing loss in the year 2000. *World Heal. Organ.* 1–30 (2000).

19. Atik, A. Pathophysiology and Treatment of Tinnitus: An Elusive Disease. *Indian Journal of Otolaryngology and Head and Neck Surgery* **66**, 1–5 (2014).

20. Coles, R. R. A. Epidemiology of tinnitus: (1) Prevalence. *J. Laryngol. Rhinol. Otol.* **98**, 7–15 (1984).

21. Axelsson, A. & Ringdahl, A. Tinnitus-A study of its prevalence and characteristics. *Br. J. Audiol.* **23**, 53–62 (1989).

22. Maas, I. L. *et al.* Genetic susceptibility to bilateral tinnitus in a Swedish twin cohort. *Genet. Med.* **19**, 1007–1012 (2017).

23. Vona, B., Nanda, I., Shehata-Dieler, W. & Haaf, T. Genetics of tinnitus: Still in its infancy. *Frontiers in Neuroscience* **11**, (2017).

24. Ivarsdottir, E. V. *et al.* Sequence variation at ANAPC1 accounts for 24% of the variability in corneal endothelial cell density. *Nat. Commun.* **10**, (2019).

25. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).

26. Hietanen, A. *et al.* Hearing among 75-year-old people in three Nordic localities: A comparative study. *Int. J. Audiol.* **44**, 500–508 (2005).

27. Roth, T. N., Hanebuth, D. & Probst, R. Prevalence of age-related hearing loss in Europe: A review. *European Archives of Oto-Rhino-Laryngology* **268**, 1101–1107 (2011).

28. Pearson, J. D. *et al.* Gender differences in a longitudinal study of age-associated hearing loss. *J. Acoust. Soc. Am.* **97**, 1196 (1995).

29. Barrenäs, M. L., Bratthall, Å. ̊ & Dahlgren, J. The association between short stature and sensorineural hearing loss. *Hear. Res.* **205**, 123–130 (2005).

30. Welch, D. & Dawes, P. J. D. Childhood hearing is associated with growth rates in infancy and adolescence. *Pediatr. Res.* **62**, 495–498 (2007).

31. Barrenäs, M. L., Jonsson, B., Tuvemo, T., Hellström, P. A. & Lundgren, M. High risk of sensorineural hearing loss in men born small for gestational age with and without obesity or height catch-up growth: A prospective longitudinal register study on birth size in 245,000 swedish conscripts. *J. Clin. Endocrinol. Metab.* **90**, 4452–4456 (2005).

32. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).

33. Wesdorp, M. *et al.* Broadening the phenotype of DFNB28: Mutations in TRIOBP are associated with moderate, stable hereditary hearing impairment. *Hear. Res.* **347**, 56–62 (2017).

34. Pollak, A. *et al.* Whole exome sequencing identifies TRIOBP pathogenic variants as a cause of post-lingual bilateral moderate-to-severe sensorineural hearing loss. *BMC Med. Genet.* **18**, 1–9 (2017).

35. Battelino, S., Klancar, G., Kovac, J., Battelino, T. & Trebusak Podkrajsek, K.

TMPRSS3 mutations in autosomal recessive nonsyndromic hearing loss. *Eur. Arch. Oto-Rhino-Laryngology* (2016). doi:10.1007/s00405-015-3671-0

36. Shen, J., Scheffer, D. I., Kwan, K. Y. & Corey, D. P. SHIELD: An integrative gene expression database for inner ear research. *Database* **2015**, (2015).

37. Hickox, A. E. *et al.* Global Analysis of Protein Expression of Inner Ear Hair Cells. *J. Neurosci.* **37**, 1320–1339 (2017).

38. Scheffer, D. I., Shen, J., Corey, D. P. & Chen, Z.-Y. Gene Expression by Mouse Inner Ear Hair Cells during Development. *J. Neurosci.* **35**, 6366–6380 (2015).

39. Santos, R. L. P. *et al.* DFNB68, a novel autosomal recessive non-syndromic hearing impairment locus at chromosomal region 19p13.2. *Hum. Genet.* **120**, 85–92 (2006).

40. Bonsch, D. *et al.* A new locus for an autosomal dominant, non-syndromic hearing impairment (DFNA57) located on chromosome 19p13.2 and overlapping with DFNB15. *HNO* **56**, 177–182 (2008).

41. Cruickshanks, K. J., Zhan, W. & Zhong, W. Epidemiology of Age-Related Hearing Impairment. in 259–274 (2010). doi:10.1007/978-1-4419-0993-0_9

42. McMahon, C. M. *et al.* The Need for Improved Detection and Management of Adult-Onset Hearing Loss in Australia. *Int. J. Otolaryngol.* **2013**, 1–7 (2013).

43. MORTON, N. E. Genetic Epidemiology of Hearing Impairment. *Ann. N. Y. Acad. Sci.* **630**, 16–31 (1991).

44. Wei, Q. *et al.* Transition fibre protein FBF1 is required for the ciliary entry of assembled intraflagellar transport complexes. *Nat. Commun.* (2013). doi:10.1038/ncomms3750

45. Tanos, B. E. *et al.* Centriole distal appendages promote membrane docking, leading to cilia initiation. *Genes Dev.* (2013). doi:10.1101/gad.207043.112

46. Grillet, N. *et al.* Mutations in LOXHD1, an Evolutionarily Conserved Stereociliary Protein, Disrupt Hair Cell Function in Mice and Cause Progressive Hearing Loss in Humans. *Am. J. Hum. Genet.* **85**, 328–337 (2009).

47. Langguth, B., Kreuzer, P. M., Kleinjung, T. & De Ridder, D. Tinnitus: Causes and clinical management. *The Lancet Neurology* **12**, 920–930 (2013).

48. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, e1001779 (2015).

49. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* https://doi.org/10.1101/166298 (2017). doi:10.1101/166298

50. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

51. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* (2015). doi:10.1038/ng.3247

52. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).

53. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).

54. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 1–8 (2019).

55. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).

56. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study

in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).

57. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, 1–7 (2017).

58. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

59. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

60. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

61. Hout, C. V. Van *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv* 572347 (2019). doi:10.1101/572347

62. Hastie, T. & Tibshirani, R. Generalized Additive Models. *Statistical Science* **1**, 297–310 (1986).

63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

64. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).

65. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

66. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

67. Oskarsson, G. R. *et al.* A truncating mutation in EPOR leads to hypo-responsiveness to erythropoietin with normal haemoglobin. *Commun. Biol.* **1**, (2018).

68. Styrkarsdottir, U. *et al.* Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. *Nature Genetics* **50**, 1681–1687 (2018).

Tables

Table 1. Summary of audiometric measures from the DHS dataset (N=11,484). For each frequency, the mean, standard deviation (SD) and range of the hearing thresholds is shown. The prevalence of mild, moderate, severe and profound hearing impairment is shown for each frequency. The effect of age in SD, sex given for women and height in SD on ARHI (PTA>25 dB HL) and the corresponding P-values are shown for each frequency.

				Hearing impairment prevalence (%)				**Effect of age and sex on ARHI**				**Effect of age, sex and height on ARHI**					
kHz	**Mean**	**SD**	**Range**	**Mild**	**Moderate**	**Severe**	**Profound**	**Age effect**	**P-value**	**Sex effect**	**P-value**	**Age effect**	**P-value**	**Sex effect**	**P-value**	**Height effect**	**P-value**
0.5	22.1	6.3	20-100	11.0%	2.5%	0.5%	0.1%	2.90	6.1×10^{-191}	1.52	2.4×10^{-10}	2.76	1.2×10^{-162}	1.09	0.37	0.80	1.1×10^{-6}
1	23.2	7.7	20-100	15.6%	4.3%	0.7%	0.1%	2.80	5.4×10^{-232}	1.21	6.9×10^{-4}	2.70	1.3×10^{-202}	0.93	0.36	0.83	7.4×10^{-6}
2	26.1	11.1	20-100	27.8%	9.5%	2.2%	0.3%	3.50	$<1\times10^{-300}$	0.78	2.6×10^{-7}	3.43	$<1\times10^{-300}$	0.67	6.3×10^{-9}	0.90	1.7×10^{-3}
4	33.4	16.7	20-100	49.1%	26.6%	8.7%	1.4%	5.50	$<1\times10^{-300}$	0.29	4.2×10^{-139}	5.44	$<1\times10^{-300}$	0.27	1.9×10^{-81}	0.93	0.038
6	32.8	16.8	20-100	45.8%	25.1%	8.8%	1.5%	5.32	$<1\times10^{-300}$	0.37	3.1×10^{-96}	5.28	$<1\times10^{-300}$	0.35	1.3×10^{-53}	0.96	0.21
8	34.2	18.7	20-90	45.5%	28.9%	12.2%	2.4%	8.66	$<1\times10^{-300}$	0.46	6.4×10^{-52}	8.60	$<1\times10^{-300}$	0.43	3.5×10^{-30}	0.96	0.21

kHz = kilohertz

Table 2. Association of sequence variants with ARHI. The table lists the 21 novel variants identified in the GWAS meta-analysis on ARHI. Gene names marked with * are novel hearing loci. For intergenic variants, the nearest genes are reported in brackets.

Model	P-value	OR	rs name	Chrom	Position	EA	OA	Gene	Variant annotation	EAF Ice (%)	EAF UK (%)
A	5.7×10^{-27}	4.20	-	17	75927880			*FBF1**	CDS tdup	0.22	0
A	8.0×10^{-14}	1.32	rs146694394	6	158071628	T	C	*SYNJ2*	missense	0.37	0.46
A	1.8×10^{-13}	1.28	rs141952919	7	103421378	G	A	*SLC26A5*	missense	0.33	0.52
A	5.9×10^{-13}	6.59	rs761934676	16	2497068	G	A	*TBC1D24*	missense	0	0.01
A	5.1×10^{-11}	1.08	rs113784020	11	118689022	T	C	*[PHLDB1]*	intergenic	3.75	4.27
A	2.5×10^{-10}	1.92	rs749405486	22	50549067	A	AG	*KLHDC7B*	frameshift	0.00	0.06
A	4.9×10^{-13}	0.95	rs72622588	3	182285702	T	G	*[FLJ46066]*	intergenic	10.73	10.81
A	6.5×10^{-10}	1.20	rs143796236	17	81528943	T	C	*FSCN2**	missense	0.21	0.74
A	7.6×10^{-10}	1.03	rs13171669	5	149221680	G	A	*ABLIM3*	intron	42.92	42.34
A	1.2×10^{-9}	1.03	rs3014246	1	45620405	C	T	*CCDC17*	missense	27.01	29.64
A	1.6×10^{-9}	1.03	rs920701	13	75842965	C	T	*LMO7*	intron	34.91	36.67
A	3.9×10^{-9}	0.97	rs11881070	19	2389142	T	C	*TMPRSS9*	upstream gene	30.15	28.79
A	4.9×10^{-8}	1.81	rs764272881	20	3228565	G	A	*SLC4A11*	missense	0.45	0
A	4.1×10^{-8}	71.17	rs765488721	17	61403195	T	C	*TBX2**	stop gained	0.01	0
A	8.3×10^{-8}	1.49	rs727503493	21	42389042	T	TG	*TMPRSS3*	frameshift	0.22	0.07
R	1.7×10^{-22}	3.65	rs118174674	18	46557437	T	C	*LOXHD1*	missense	2.95	1.99
R	3.2×10^{-11}	1.05	rs9394952	6	43433367	G	A	*ABCC10*	splice region	49.99	48.13
R	7.7×10^{-11}	2.35	rs12784122	10	80649861	A	G	*SH2D4B**	downstream gene	2.66	2.29
R	3.5×10^{-10}	17.32	rs139123090	10	126459169	A	G	*C10orf90**	missense	0.89	0.47
R	3.2×10^{-9}	0.94	rs557563970	1	117960109	CGT	C	*WDR3*	3 prime UTR	21.76	41.74
R	1.2×10^{-8}	4.79	rs74543584	11	118262596	A	T	*MPZL2*	missense	1.47	0.83

OR=Odds ratio, Chrom=Chromosome, EA=Effect Allele, OA=Other allele, EAF=Effect allele frequency, Ice=Iceland, A=Additive model, R=Recessive model

Figures

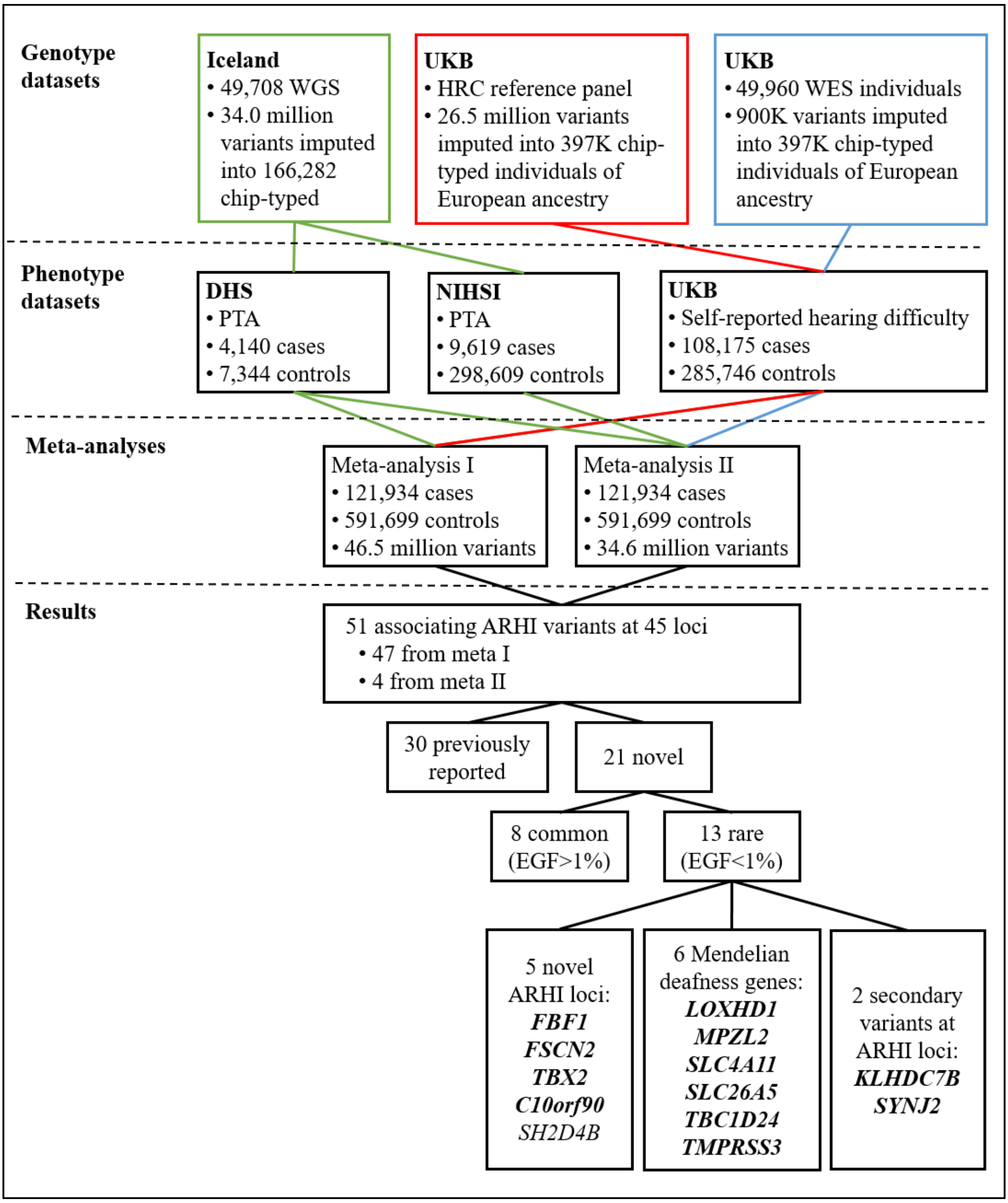

Figure 1. Study design and summary of results. The UK Biobank (UKB) GWAS was performed with two genotype datasets marked in red and blue. The gene names in bold are the loci where the association is represented by a coding variant.

WGS = Whole genome sequenced, HRC = Haplotype Reference Consortium, WES = Whole exome sequenced, PTA = Pure tone average, DHS = deCODE health study, NIHSI = National Institute of Hearing and Speech in Iceland, EGF = Expected genotype frequency

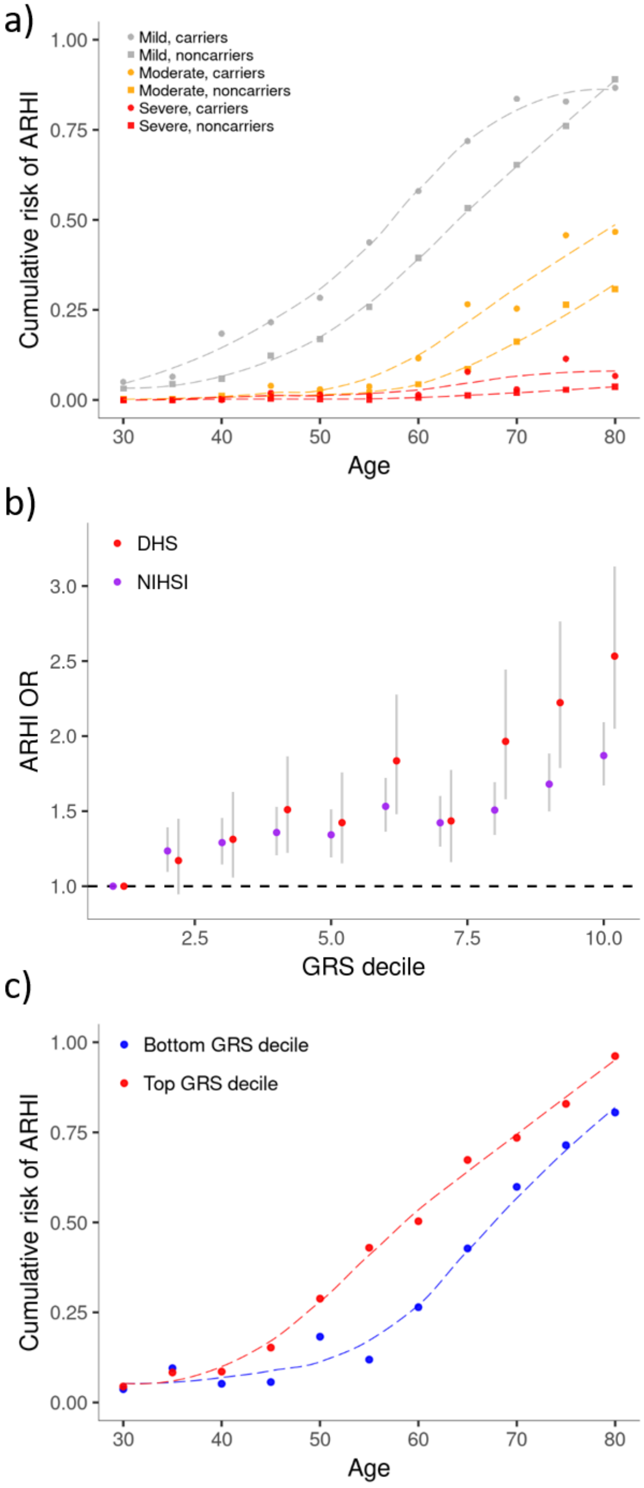
a)
Cumulative risk of ARHI
1.00
0.75
0.50
0.25
0.00
● Mild, carriers
■ Mild, noncarriers
● Moderate, carriers
■ Moderate, noncarriers
● Severe, carriers
■ Severe, noncarriers
30
40
50
60
70
80
Age
b)
ARHI OR
3.0
2.5
2.0
1.5
1.0
● DHS
● NIHSI
2.5
5.0
7.5
10.0
GRS decile
c)
Cumulative risk of ARHI
1.00
0.75
0.50
0.25
0.00
● Bottom GRS decile
● Top GRS decile
30
40
50
60
70
80
Age

Figure 2. ARHI risk for rare variants and a common variant GRS. a) The cumulative risk of mild, moderate and severe hearing impairment in the DHS dataset among the 4.9% of subjects that are carriers of any of the 16 rare ARHI variants (dots) and the 95% that are not carriers (squares). **b)** The ORs for ARHI for each GRS decile in the DHS and NIHSI datasets compared to the bottom decile. **c)** The cumulative risk of ARHI among subjects in the DHS dataset in the bottom GRS decile in blue and the top GRS decile in red.

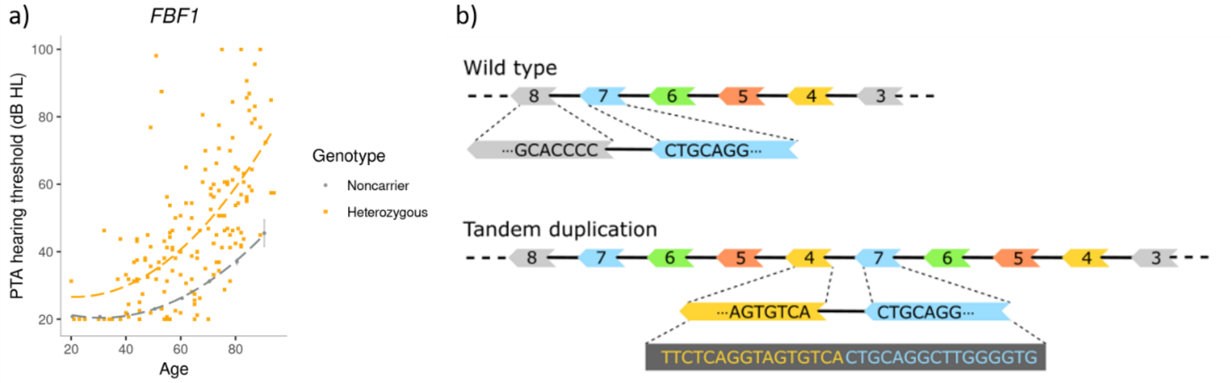

Figure 3. A tandem duplication in *FBF1* associates with ARHI. a) The PTA hearing threshold of the heterozygous carriers of the tandem duplication in *FBF1* in DHS and NIHSI datasets are indicated by yellow squares and the average PTA hearing thresholds of non-carriers in the DHS dataset are represented with grey dots. **b)** The exon structure of the wild type and the tandem duplication variant in transcript ENST00000586717.5 of *FBF1* on the reverse strand of chromosome 17. The exons of the transcripts are numbered, and solid black lines represent splicing between exons. The tandem duplication creates a longer transcript with extra sets of exons 4 to 7 that leads to novel splice junction starting at the end of exon 7 and splicing into the beginning of exon 4. The adjacent sequences of the exons are shown, and together they form the 32bp sequence used for identification of the novel junction.

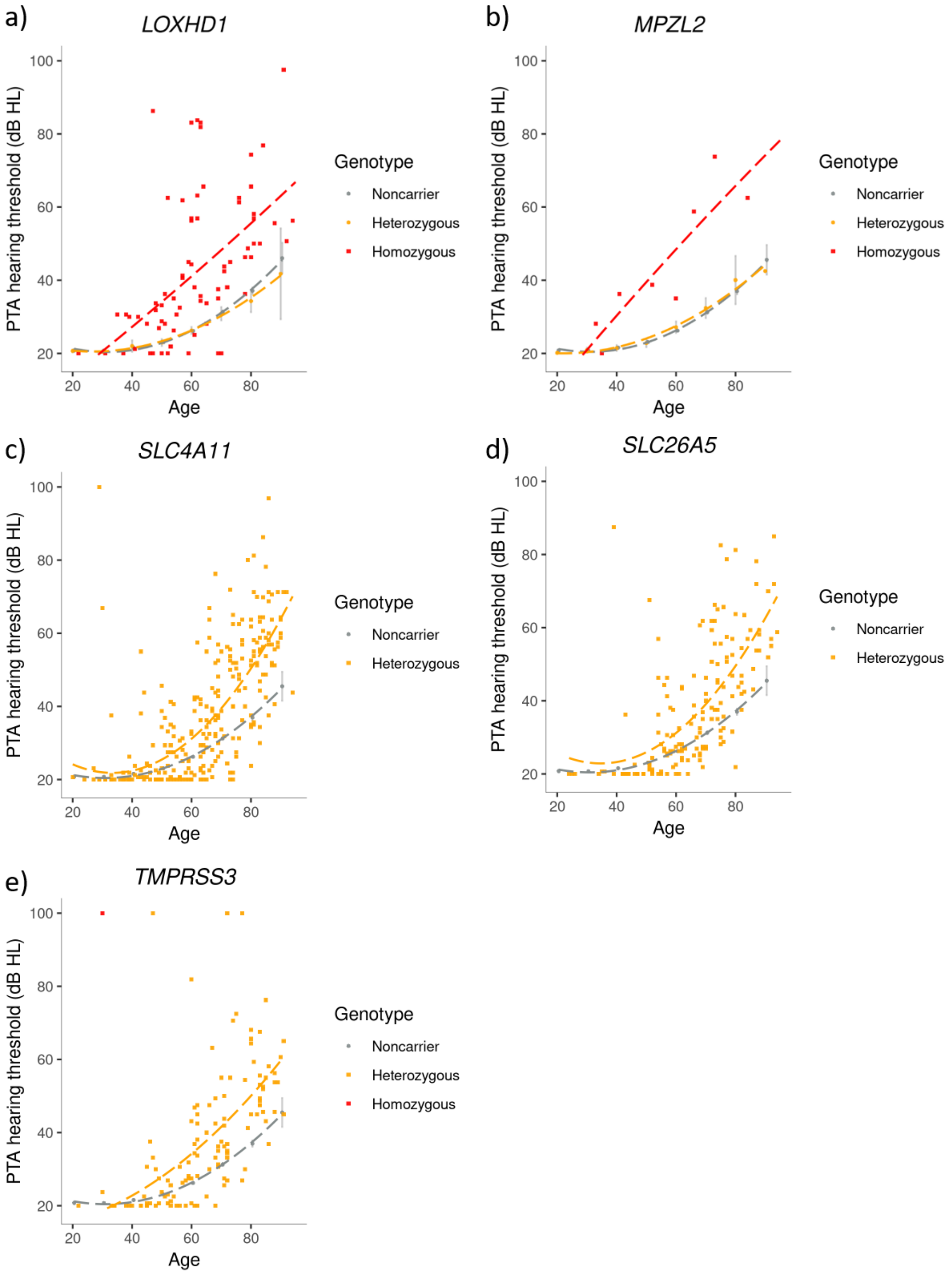

Figure 4. Changes in PTA hearing thresholds by age for carriers of rare ARHI variants in Mendelian deafness genes. Effects of variants in **a)** *LOXHD1*, **b)** *MPZL2*, **c)** *SLC4A11*, **d)** *SLC26A5* and **e)** *TMPRSS3* are shown. In figure a and b, the average PTA in the DHS dataset

are represented with grey dots for non-carriers and orange dots for heterozygotes and the PTA hearing thresholds of the homozygous carriers in DHS and NIHSI datasets are represented with red squares. In figures c, d and e, the average PTA of non-carriers in the DHS dataset are represented with grey dots and the PTA hearing threshold of the heterozygous carriers in DHS and NIHSI datasets by yellow squares. A figure for *TBC1D24* is not included because the variant was detected in UKB dataset only and therefore audiometric measures are not available for the carriers.

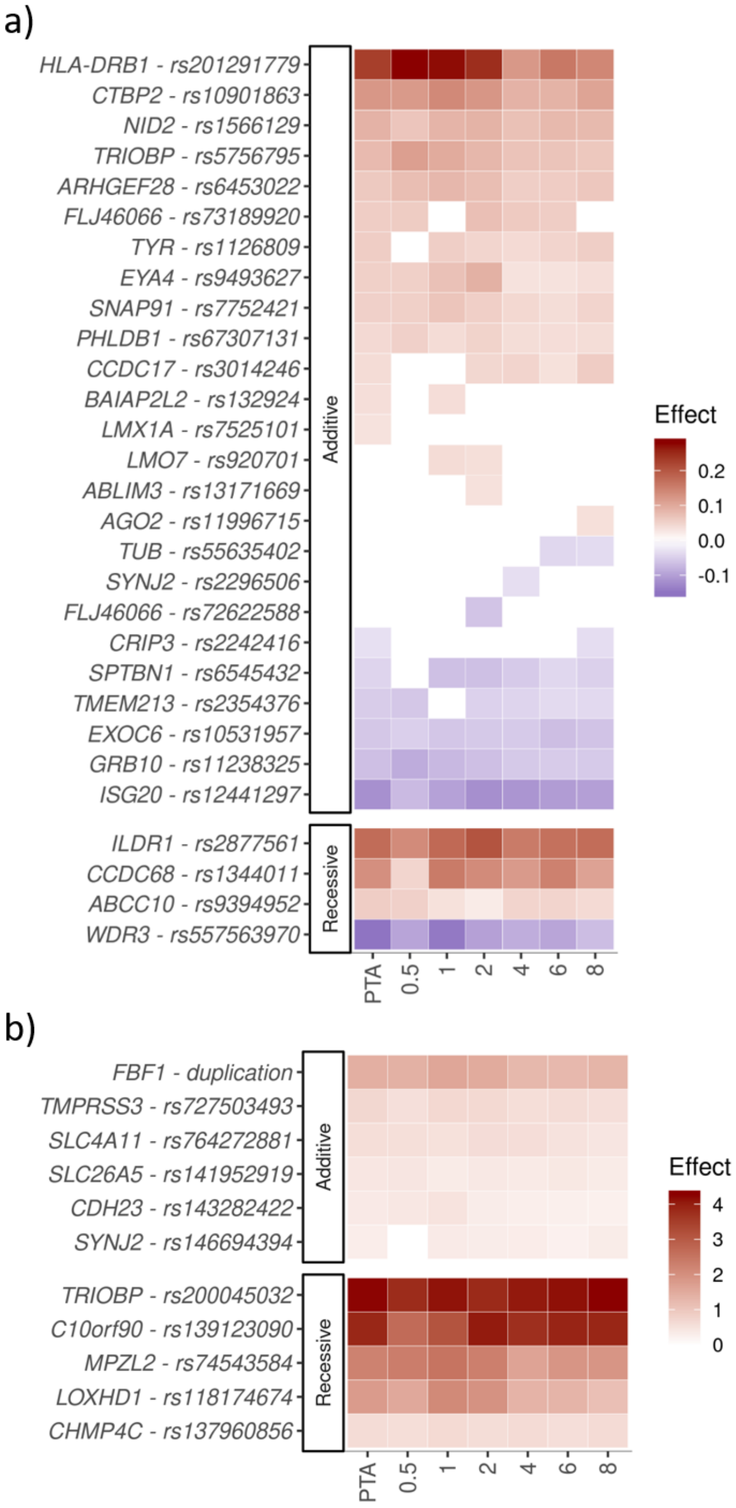
a)
HLA-DRB1 - rs201291779
CTBP2 - rs10901863
NID2 - rs1566129
TRIOBP - rs5756795
ARHGEF28 - rs6453022
FLJ46066 - rs73189920
TYR - rs1126809
EYA4 - rs9493627
SNAP91 - rs7752421
PHLDB1 - rs67307131
CCDC17 - rs3014246
BAIAP2L2 - rs132924
LMX1A - rs7525101
LMO7 - rs920701
ABLIM3 - rs13171669
AGO2 - rs11996715
TUB - rs55635402
SYNJ2 - rs2296506
FLJ46066 - rs72622588
CRIP3 - rs2242416
SPTBN1 - rs6545432
TMEM213 - rs2354376
EXOC6 - rs10531957
GRB10 - rs11238325
ISG20 - rs12441297
Additive
ILDR1 - rs2877561
CCDC68 - rs1344011
ABCC10 - rs9394952
WDR3 - rs557563970
Recessive
PTA
0.5
1
2
4
6
8
Effect
0.2
0.1
0.0
-0.1
b)
FBF1 - duplication
TMPRSS3 - rs727503493
SLC4A11 - rs764272881
SLC26A5 - rs141952919
CDH23 - rs143282422
SYNJ2 - rs146694394
Additive
TRIOBP - rs200045032
C10orf90 - rs139123090
MPZL2 - rs74543584
LOXHD1 - rs118174674
CHMP4C - rs137960856
Recessive
PTA
0.5
1
2
4
6
8
Effect
4
3
2
1
0

Figure 5. Effects of the ARHI variants on ARHI per frequency. Each row shows the estimated effect of the minor allele on ARHI for PTA, the average of 0.5,1,2 and 4kHz, and separately for each frequency, 0.5, 1, 2, 4, 6, and 8 kHz, for **a)** common variants with EGF>1% and **b)** rare variants with EGF<1%. The effect is shown only for associations with P-value<0.05. Red color represents increased risk of ARHI and blue color represents decreased risk.

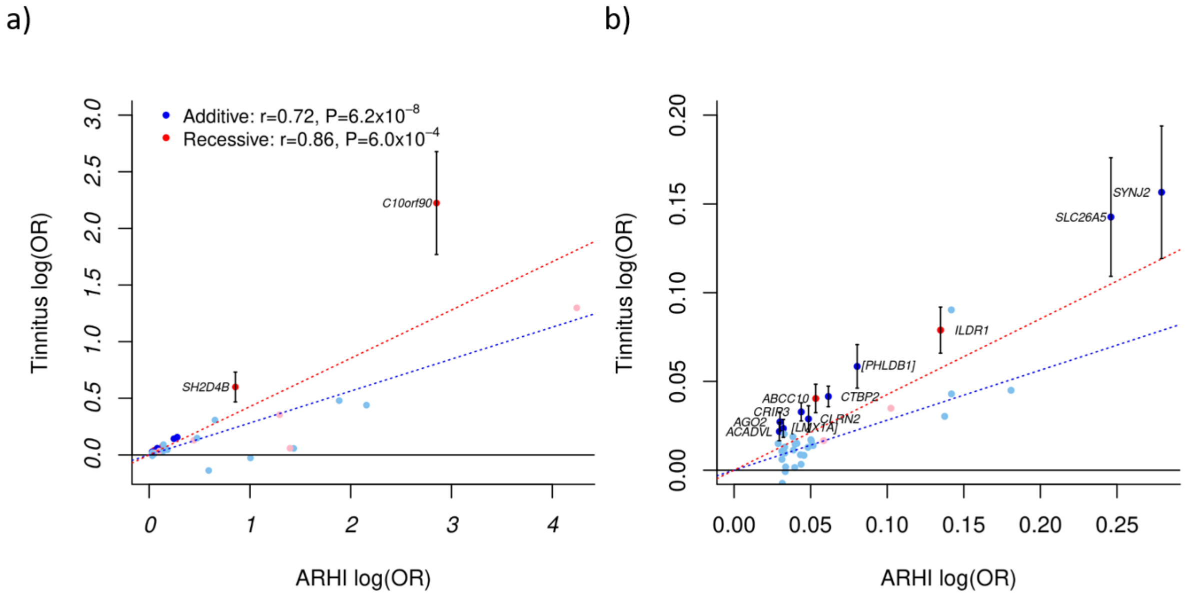

Figure 6. Effect of the ARHI variants on tinnitus. The effect of the ARHI variants on ARHI is plotted against their effect on tinnitus for **a)** all ARHI variants and **b)** zoomed-in on variants from a) with ARHI OR <1.35. ARHI variants detected under the additive model were tested for tinnitus using the additive model (colored blue) and ARHI variants detected under the recessive model were tested for tinnitus using the recessive model (colored red). Variants that affect tinnitus, controlling the false discovery rate at 0.05, are plotted with darker color and labelled with their corresponding gene. All effects are shown for the ARHI risk increasing allele. Error bars represent 95% confidence intervals. The dotted lines represent results from a weighted linear regression using MAF(1-MAF) as weights, red for recessive

variants and blue for additive, and the weighted correlation coefficients (r) and the corresponding P-values are shown in figure a.