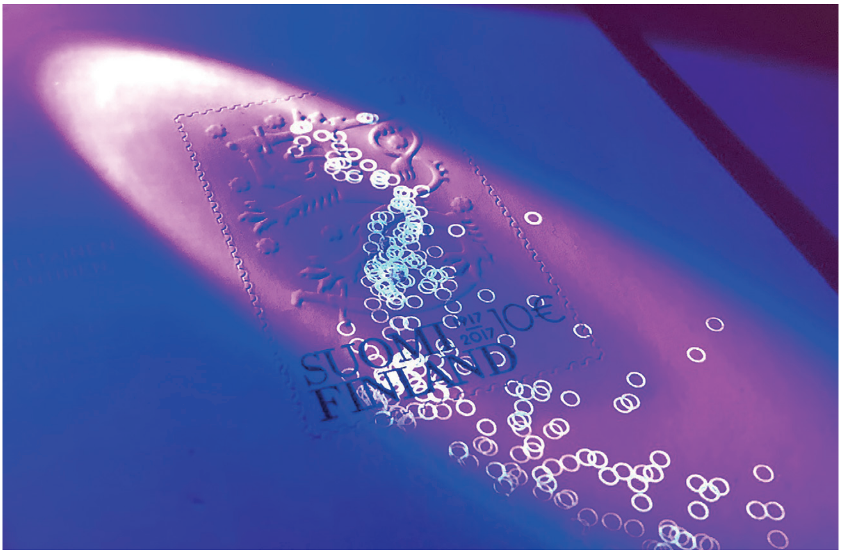
DISSERTATIONES SCHOLAE DOCTORALIS AD SANITATEM INVESTIGANDAM
UNIVERSITATIS HELSINKIENSIS

SINI KERMINEN

FINE-SCALE GENETIC STRUCTURE AND POLYGENIC SCORES IN FINLAND



INSTITUTE FOR MOLECULAR MEDICINE FINLAND (FIMM)
HELSINKI INSTITUTE OF LIFE SCIENCE (HiLIFE)
FACULTY OF MEDICINE
DOCTORAL PROGRAMME IN POPULATION HEALTH
UNIVERSITY OF HELSINKI

FINE-SCALE GENETIC STRUCTURE AND POLYGENIC SCORES IN FINLAND

SINI KERMINEN

Institute for Molecular Medicine Finland (FIMM)
Helsinki Institute of Life Science (HiLIFE)
Faculty of Medicine
Doctoral Programme in Population Health
Doctoral School in Health Sciences
University of Helsinki
Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Lecture Hall 1, Haartman Institute, on the 29th of January 2021, at 14.00.

Helsinki 2021

UNIVERSITY OF HELSINKI

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

ISBN 978-951-51-6907-5 (Print)
ISBN 978-951-51-6908-2 (PDF)
ISSN 2342-3161 (Print)
ISSN 2342-317X (Online)

Hansaprint Oy
Turku 2021

Cover image:
A special stamp ‘Suomen vaakuna 2017’ reveals the genetic structure of Finland under UV light. The stamp is designed by Pekka Piippo and published by Posti. Photo by Sini Kerminen.

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

Supervisors

Dr Matti Pirinen

Institute for Molecular Medicine Finland (FIMM)
Helsinki Institute of Life Science (HiLIFE)
University of Helsinki, Finland

Department of Public Health
Faculty of Medicine
University of Helsinki, Finland

Department of Mathematics and Statistics
University of Helsinki, Finland

Prof Samuli Ripatti

Institute for Molecular Medicine Finland (FIMM)
Helsinki Institute of Life Science (HiLIFE)
University of Helsinki, Finland

Department of Public Health
Faculty of Medicine
University of Helsinki, Finland

Broad Institute of MIT and Harvard
Cambridge, MA, USA

Reviewers

Prof Outi Savolainen

Department of Ecology and Genetics
Faculty of Natural Science
University of Oulu, Finland

Dr Kristiina Tambets

Estonian Biocenter
Institute of Genomics
University of Tartu, Estonia

Opponent

Dr Tuuli Lappalainen

Department of Systems Biology
Columbia University

New York Genome Center,
New York, NY, USA

TIIVISTELMÄ

Populaatiogenetiikka on nykyisin olennainen osa ihmisen alkuperän ja historian sekä tautien geneettistä tutkimusta. Erillään toisistaan elävien populaatioiden välillä havaitaan geneettisiä eroja, ja tätä populaatioiden välistä vaihtelua kutsutaan geneettiseksi rakenteeksi. Tutkimalla geneettistä rakennetta eri puolilla maailmaa on esimerkiksi pystytty tarkentamaan nykyihmisen levittäytymisreittejä Afrikan mantereelta muualle maailmaan viimeisen 100,000 vuoden aikana. Populaatiogenetiikan modernit menetelmät ovat myös mahdollistaneet geneettisen rakenteen ja populaatiohistorian yksityiskohtaisen analysoinnin yksittäisten maiden sisällä, mutta näitä menetelmiä ei ole vielä laajasti hyödynnetty eristäytyneissä populaatioissa kuten Suomessa.

Lääketieteellisessä genetiikassa tautien geneettistä taustaa analysoidaan rutiininomaisesti perimänlaajuisissa assosiaatiotutkimuksissa (GWAS). Näissä tutkimuksissa on tärkeää kontrolloida aineiston geneettinen rakenne kunnolla, jotta voidaan luotettavasti erottaa tautiin liittyvä geneettinen vaihtelu populaatioiden geneettiseen rakenteeseen liittyvästä yleisestä vaihtelusta. GWAS-tutkimusten tuloksia käytetään tieteellisessä tutkimuksessa ennustamaan yksilöiden geneettistä sairastumisriskiä polygeensellä riskiarvolla, joka summaa yhteen usean perimänkohdan arvioidun geneettisen riskin. Summaamisesta johtuen pienikin geneettisestä rakenteesta johtuva harha GWAS-tuloksissa voi aiheuttaa merkittävää harhaa polygeenisiin riskiarvoihin, mikä puolestaan voi johtaa virheellisiin johtopäätöksiin erityisesti populaatioiden välisissä vertailuissa. Näin ollen geneettisen rakenteen tunteminen ja sen rooli polygeenisten riskiarvojen koostamisessa onkin erityisen tärkeää, jotta voimme ymmärtää sekä polygeenisten riskiarvojen hyödyt että rajoitteet akateemisessa tutkimuksessa ja mahdollisissa tulevaisuuden terveydenhuollon sovelluksissa.

Tässä väitöskirjatutkimuksessa tutkittiin suomalaisten geneettistä hienorakennetta ja sen roolia polygeenisten riskiarvojen maantieteellisessä jakautumisessa Suomessa. Väitöskirjan ensimmäinen osa laajensi ymmärrystä Suomen geneettisestä rakenteesta määrittämällä maantieteellisen rajan Suomen geneettiselle pääjaolle Itä- ja Länsi-Suomen välillä ja tunnistamalla 17 ennen näkemätöntä geneettistä hienopopulaatiota. Hienopopulaatioiden havaittiin olevan maantieteellisesti keskittyneitä ja noudattelevan Suomen murrealueita. Toisessa osassa hyödynnettiin aiempia tuloksia muodostamalla hienorakenteen pohjalta vertailuryhmät yksilön geneettisen alkuperäprofiilin määrittämiselle Suomen sisällä. Määrittämällä geneettinen alkuperäprofiili joukolle 1923 ja 1987 välillä syntyneitä yksilöitä, tutkimus onnistui kartoittamaan vuosittaisia muutoksia Suomen geneettisessä hienorakenteessa 12 alueella. Vuosittaiset profiilit vastasivat hyvin 1939–1945 sotatapahtumien seurauksena käynnistyneitä karjalaisten evakkojen muuttoliikkeitä. Kolmas osa arvioi edellä mainitun geneettisen itä-länsi jaon roolia viiden monitekijäisen taudin (sepelvaltimotauti, nivelreuma, skitsofrenia, haavainen koliitti ja Crohnin tauti) ja kolmen mitattavan ominaisuuden (pituus, painoindeksi ja lantio-vyötärö-suhde) geneettisen riskin maantieteellisen jakauman taustalla polygeenisiä riskiarvoja käyttäen. Tutkimus osoitti, että useimmat polygeeniset riskiarvot, joissa maantieteellisiä eroja havaittiin, heijastelivat geneettistä jakoa Itä- ja Länsi-Suomen välillä, mutta sisälsivät myös geneettiseen rakenteeseen liittyvää tilastollista virhettä. Tutkimus osoitti, että polygeeniset riskiarvot ovat alttiita geneettiseen rakenteeseen liittyvälle harhalle myös suhteellisen samankaltaisissa populaatioissa ja että yhteyttä populaatioiden geneettisen vaihtelun ja alueellisten sairastuvuuserojen välillä on vaikea osoittaa.

Kokonaisuutena tämän väitöskirjan tulokset päivittivät ymmärryksen Suomen yksityiskohtaisesta geneettisestä rakenteesta ja sen muutoksista vastaamaan nykyaikaisen geneettisen tutkimuksen tarpeita, ja havainnollisti sekä tiedeyhteisölle että suurelle yleisölle geneettisen rakenteen tuntemuksen merkityksen niin populaatiohistorian kuin polygeenisten riskiarvojen tutkimuksessa.

ABSTRACT

Population genetics is today an essential part of the studies of human origin and history, as well as of the studies of disease genetics. Populations living apart from each other exhibit genetic variation and this variation between populations is called genetic structure. By studying the genetic structure around the world, it has become possible, for example, to elaborate on the migration patterns of modern humans from the African continent to the rest of the world during the last 100,000 years. In addition, the modern methods of population genetics have enabled a very detailed analysis of genetic structure and population history within single countries, but these methods have not yet been widely utilized in isolated populations such as in Finland.

In medical genetics, the genetic background of diseases is routinely analyzed with a genome-wide association study (GWAS). In these studies, it is important to control for the genetic structure of the data appropriately, so that the genetic variation associated with the disease can be reliably distinguished from the general genetic variation associated with the genetic structure. The results of GWAS are used in research to predict the genetic disease risk of an individual using a polygenic score that summarizes the estimated genetic risk over multiple sites of the genome. Because of this summation, even a tiny bias in the GWAS results, due to the genetic structure, can lead to a significant bias in polygenic scores, which, in turn, can lead to incorrect conclusions especially in comparisons between populations. Therefore, understanding the genetic structure and its role in building polygenic scores is exceptionally important to properly understand both the benefits and limitations of polygenic scores in academic research and in future health-care applications.

This doctoral thesis examined the fine-scale genetic structure and its role in the geographic distribution of polygenic scores in Finland. The first part of the thesis expanded the understanding of the genetic structure of Finland by determining the geographic border for the major genetic split between East and West Finland, and by identifying 17 previously unreported genetic fine-scale populations. The fine-scale populations were observed to be geographically clustered and to follow Finnish dialect regions. The second part of the thesis utilized the earlier results by building, based on the fine-scale genetic structure, reference groups to estimate the genetic ancestry profile of an individual within Finland. By estimating the genetic ancestry profiles for a set of individuals born between 1923 and 1987, this second study was able to map annual changes in the fine-scale genetic structure within 12 regions. The annual profiles matched well with the migration patterns of Karelian evacuees who were displaced by the war events between 1939 and 1945. The third part of the thesis assessed the role of the genetic split between East and West in the geographic distribution of the genetic risk of five complex diseases (coronary artery disease, rheumatoid arthritis, schizophrenia, ulcerative colitis, and Crohn's disease) and three quantitative traits (height, body mass index, and waist-hip ratio) using polygenic scores. The third study demonstrated that most of the polygenic scores, that did show geographic variation, mirrored the genetic split between East and West Finland but also revealed bias associated with the genetic structure in Finland. This final study thus demonstrated two main points: first that polygenic scores are susceptible to genetic structure-related biases, even within a relatively homogeneous populations; and second that it is challenging to link population-genetic variation to geographic variation in disease incidence with the current methods.

Overall, the results of this doctoral thesis update the current understanding of fine-scale genetic structure in Finland and its changes to meet the needs of modern genetics research. It also demonstrates, both for the scientific community and for the general public, the importance of understanding the genetic structure in the study of both population history and polygenic scores.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following three publications and they are referred to in the text by their Roman numerals:

I **Kerminen S.**, Havulinna A. S., Hellenthal G., Martin A. R., Sarin A-P., Perola M., Palotie A., Salomaa V., Daly M. J., Ripatti S. and Pirinen M., Fine-Scale Genetic Structure in Finland, G3: Genes, Genomes, Genetics (2017) vol. 7, no. 10: 3459-3468.

II **Kerminen S.**, Cerioli N., Pacauskas D., Havulinna A. S., Perola M., Jousilahti P., Salomaa V., Daly M. J., Vyas R., Ripatti S. and Pirinen M., Changes in the Fine-Scale Genetic Structure of Finland Through the 20th Century, submitted.

III **Kerminen S**., Martin A. R., Koskela J., Ruotsalainen S. E., Havulinna A. S., Surakka I., Palotie A., Perola M., Salomaa V., Daly M. J., Ripatti S. and Pirinen M., Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland, American Journal of Human Genetics (2019) vol. 104, no. 6: 1169-1181.

ABBREVIATIONS

A	Adenine
aDNA	Ancient DNA
BMI	Body mass index
bp	Base pair
BP	Before present
C	Cytosine
CAD	Coronary artery disease
CD	Crohn's disease
CE	Common era
cM	centimorgan
DNA	Deoxyribonucleic acid
G	Guanine
GWAS	Genome-wide association study
HG	Height
LD	Linkage disequilibrium
MAF	Minor allele frequency
MCMC	Markov chain Monte Carlo
mt-	Mitochondrial
PCA	Principal component analysis
PS	Polygenic score
RA	Rheumatoid arthritis
Refset	Reference set
SCZ	Schizophrenia
SNP	Single nucleotide polymorphism
T	Thymine
UC	Ulcerative colitis
WHR	Waist-hip ratio
WWII	World War II

1 INTRODUCTION

Genetic information on human populations provides us with unique opportunities to explore not only the biology of a human being but also the origin and the history of our species. The field that studies the genetic variation of populations and the processes affecting it is called population genetics. The key concept of population genetics is genetic structure, which describes the patterns of genetic variation within and between populations, and it is routinely utilized in the studies of forensics, population history, and medical genetics. In forensics, the individual's genotype is contrasted with the genetic structure of the background population, while in the studies of population history, population genetic tools are used to identify genetic similarities between the studied groups. For example, by comparing the genetic information of modern humans and an ancient sample of a Neanderthal human, modern humans outside of Africa have been shown to share around 2% of their genome with Neanderthals demonstrating, that these two human subspecies have interbred[1].

In medical genetics, the last decade has seen an increase in the number and size of genome-wide association studies (GWAS) aiming to identify genetic variation underlying complex diseases. To avoid false positive results due to the confounding caused by genetic variation between populations, understanding the genetic structure has become an essential part of these studies. However, as the studies grow and examine even rarer variants, the current knowledge and methods controlling for genetic structure may not be sufficient. In addition, the recent attempts to build polygenic risk estimates from GWAS results, for potential use in health care, have shown that the current applications can crucially depend on the populations they were generated in and may not transfer to other populations[2, 3]. Therefore, a thorough understanding of the genetic structure between populations is needed.

This doctoral thesis focuses on the population of Finland. Finland has for decades participated in international genetic studies, because of the active research community, comprehensive national records and, most importantly, the population that harbors characteristics beneficial for genetic studies. Because of the population history, which includes isolation and genetic bottlenecks, some genetic variants—rare elsewhere in the world—are enriched in Finland and, therefore, are easier to identify in the Finnish population. Indeed, a biobank-scale research project, the FinnGen project (www.finngen.fi), is currently collecting and analyzing 500,000 Finnish genomes. Thus, Finnish genetic studies will most likely continue as a key part of the international genetics research also in the future.

Prior to this thesis, knowledge about the genetic structure of Finland has relied on the analyses of a few markers in the Y chromosome and mitochondrial genome, as well as on the analyses of sparse genome-wide markers. Consequently, these analyses have mainly characterized the genetic variation between East and West Finland, and there is a demand for more detailed information on the genetic structure of Finland and its role in the genetic studies of complex diseases. To update information on the genetic structure in Finland in order to meet the needs of modern genetic analyses, this thesis examines Finland's fine-scale genetic structure during the 20th century together with its connections to the polygenic scores of complex diseases. This thesis utilizes modern haplotype-based methods of chromosome painting[4] and the data of over 18,000 individuals from the National FINRISK Study, providing both a spatially and temporally detailed view on the fine-scale genetic structure in Finland.

To avoid misunderstanding, it is important to realize the results of this work cannot be used to define who is, or who is not, a Finnish individual. Nationality and genetic ancestry are separate concepts that do not define one another. Nationality and national identity are a diverse collection of legal, social, cultural, religious, physical, and linguistic characteristics that can be changed and obtained over a person's lifetime. Genetic ancestry, in turn, refers to the set of ancestors from whom a person inherits genetic material and thus cannot be changed. Similarly, genetic ancestry does not determine race or tribal identity, which are more complex concepts than can be defined by genetics alone. Additionally, it should be noted that the work in this thesis is limited to one major genetic group in Finland for which there are enough samples available for study, and the work does not consider Finnish minority groups or ancestry from other countries, even if these sources are an essential part of the current Finnish gene pool.

2 LITERATURE REVIEW

2.1 HUMAN GENOME

2.1.1 STRUCTURE OF THE GENOME

Genetic information is encoded in a macromolecule called deoxyribonucleic acid (DNA). DNA is constructed of two polynucleotide chains whose basic unit, the nucleotide, is constructed of a deoxyribose sugar, a phosphate group and a nitrogenous base. In DNA, four types of bases exist: adenine (A), thymine (T), cytosine (C) and guanine (G). The bases are linked to each other via the deoxyribose-phosphatase backbone forming a DNA sequence. This sequence preserves the genetic information and is used to build proteins in a cell. The two strands of DNA sequence are joined by bases, such that A connects with T, and C connects with G, allowing an efficient and accurate DNA replication.

The genetic material of a human being, known as the human genome, consists of over 3 billion base pairs (bp) that are arranged into protein-controlled macromolecules called chromosomes. As opposed to a haploid organism with only a single copy of a chromosome, humans are called diploids: the chromosomes exist in pairs. Humans have 22 autosomal chromosome pairs and two sex chromosomes: females have two copies of the sex chromosome X, while males have one X and one Y chromosome. In addition, the human genome includes extranuclear mitochondrial DNA (sometimes referred to as the mitochondrial genome) that exists from hundreds to hundreds of thousands of copies in each cell, depending on the cell type.

2.1.2 GENETIC VARIATION

While most of the DNA sequence is identical between humans, there exists genetic positions (loci, singular locus) where two individuals show genetic variation. Genetic variation is divided into two classes: structural variation and simple genetic variation. Structural variation is a large genetic change defined commonly as affecting over 1,000 bps[5]. This class includes large duplications and deletions, translocations and inversions, copy number variation, and transposable elements. However, the studies of population genetics usually focus on simple genetic variation and, more specifically, on single nucleotide polymorphisms (SNPs). SNPs are variants where at least two versions of a single base pair exist in a population with a sufficient frequency. The different versions of a variant are called alleles and the combination of the two alleles, which an individual carries, is called genotype. Other types of simple genetic variation are small insertions and deletions (indels), and short tandem repeats (microsatellites).

The international sequencing effort of the 1000 Genomes Project[6] has provided a comprehensive basis for understanding the genetic variation in the human genome. The project sequenced the whole genomes of more than 2,500 individuals from 26 world-wide populations and identified over 88 million simple genetic variants and 60,000 structural variants. Most of the simple genetic variants (~75%) were found to be rare with minor allele frequency (MAF) under 0.5%, and only 10% were common with MAF above 5%. They also estimated that, on average, an individual carries 4 to 5 million differences compared to the reference genome. Later, other projects have supplemented these analyses by providing information from additional populations, e.g., Simons Genome Diversity Project[7], or from a larger number of samples, e.g., the Exome Aggregation Consortium (ExAC)[8]. To date, the largest collection of publicly available sequencing information, the Genome Aggregation Database (gnomAD), has identified over 230 million genetic variants among the 141,000 samples[9]. Together these projects have demonstrated that most of the variation is found in African populations, and the least variation is found in isolated populations, such as the Finns.

While sequencing is an accurate method to identify both known and novel genetic variants in a sample, it is expensive to sequence samples on a large scale. A more cost-effective method for detecting known genetic variation is genotyping. The method is based on a genotyping chip (also known as a microarray) on which predesigned oligonucleotide probes have been attached to predefined positions[10]. The probes are designed in such a way that they include complementary DNA sequence for the variants and alleles of interest. The chip is exposed to a single stranded sample DNA that then anneals with the complementary probes. A successful annealing induces light emission (fluorescence) that gets recorded. Finally, the intensity of fluorescence signal together with its position is computationally transformed into genotype calls.

2.1.3 INHERITANCE AND RELATEDNESS

Genetic information is distributed from parents to an offspring in chromosomes via meiosis, the cell division of reproductive cells. However, genetic material does not stay intact in the process. Physical changes in base pairs, also known as mutations, usually affect only a few bases (with a probability of $1.1\text{-}1.6 \cdot 10^{-8}$ per bp in a generation[11], i.e., roughly 40 bases per individual), but the exchange of genetic material between chromosome pairs, called meiotic recombination, affects large chunks of a chromosome. Meiotic recombination is estimated to happen around 1.6 times per chromosome in a generation, although the rate differs between and within the chromosomes[12, 13]. Figure 1 illustrates the segregation of one chromosome in meiosis and demonstrates the formation of haploid daughter cells, whose genetic content differs from both of the parental genomes. In addition, Figure 1 shows that the genetic material is inherited in parts, i.e., in groups of alleles inherited from a single parent, and these parts are called haplotypes (sometimes also used to refer to the whole chromosome).

Some pieces of haplotypes are inherited through generations, as illustrated in Figure 2. In each generation, the pieces are broken down, shuffled and half of the genetic material is randomly transmitted to the next generation. This results in an offspring who has a random combination of haplotypes inherited from its ancestors. The number and length of shared haplotype chunks decrease, the further back in generations we go, in such a way that, on average, $1 / 2^g$ of individual's genome is shared with an ancestor from *g* generation back. This proportion of shared genetic information is called genetic relatedness (identity-by-descent) and is, for example, 0.5 between a parent and an offspring, and around 0.25 between a grandparent and an offspring. More specifically, the coefficient of genetic relatedness is defined as the total probability that the alleles of two individuals are identity-by-descent, i.e., they are inherited from a common ancestor within a certain timeframe. This definition allows us to estimate the relatedness between, for example, full siblings or first cousins by inferring the probability that they share the same haplotype via either both parents (siblings) or grandparents (cousins). Thus, the relatedness of siblings is around 0.5 and between cousins around 0.125. Inferring the unknown relationships between two individuals can be done by estimating relationship coefficient from the data. Traditionally, the estimates have been done

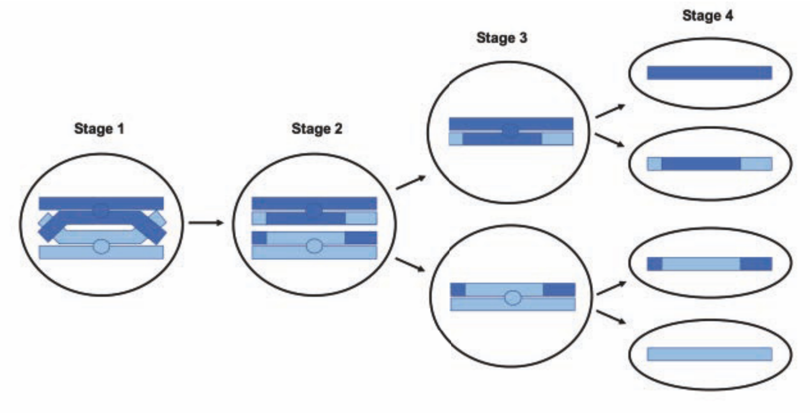


Figure 1. Schematic presentation of recombination and segregation of a chromosome in meiosis. The dark and light bars represent the homologous chromosomes. Bars of the same color represent the sister chromatids connected with centromere (filled circle). During the meiosis, the homologous chromosomes pair and recombine by crossing over (stage 1), forming sister chromatids that differ from the original chromosomes (striped) (stage 2). The homologous chromosomes are separated into newly formed cells (stage 3), and sister chromatids are further separated into haploid cells (stage 4). These haploid genomes are different from both of the parental genomes due to the recombination and the random segregation of chromosomes.

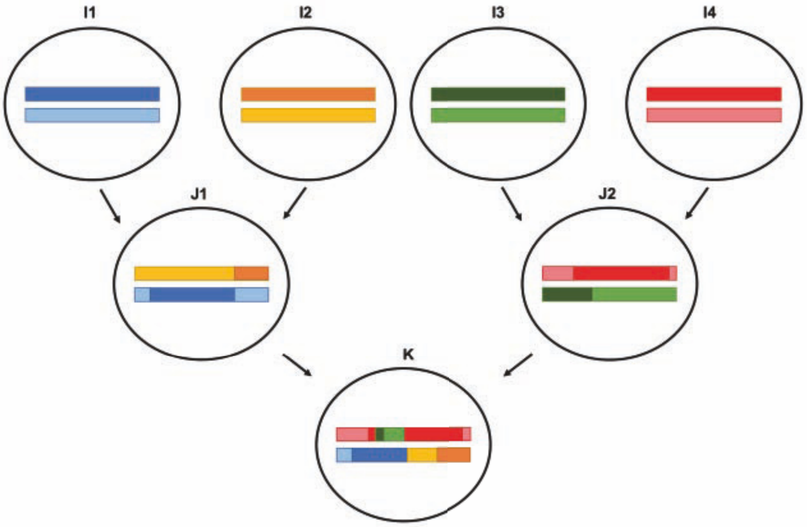


Figure 2. Schematic presentation of the breakdown and inheritance of haplotypes from grandparents (I1-I4) to parents (J1, J2), and to an offspring (K). The parent J1 inherits two chromosomes (also called haplotypes), blue and yellow, from the grandparents I1 and I2. These chromosomes are not identical to the chromosomes of grandparents but show a unique combination of them (presented with different shades of blue and yellow). Similarly, offspring K inherits a colorful mosaic of the grandparental haplotypes.

from a small number of microsatellites or SNPs[14], but the modern methods utilize genome-wide data[15, 16].

2.1.4 LINKAGE DISEQUILIBRIUM

At a population level, the genetic variants are observed to be correlated, i.e., some alleles of the nearby variants are inherited together more often than would be expected based on their allele frequencies alone. This correlation between variants is called linkage disequilibrium (LD) and it is measured between two loci, A and B (with alleles *A/a* and *B/b*), with a correlation coefficient R^2:

$$R^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b},$$

where p_{AB} is the frequency of alleles *A* and *B* observed together in the same haplotype, and p_A, p_a, p_B and p_b are the corresponding marginal allele frequencies. When R^2 is 0, the loci are in linkage equilibrium, and when R^2 is 1, the loci are in complete LD.

Reflecting the pattern of inheritance, the early studies on LD suggested that the regions of high-LD form block-like structures (haplotype blocks) that contain a limited number of haplotypes, and this information could be utilized to identify disease associated haplotypes at a population level[17-19]. However, the information on the LD pattern and genomic variation of the human genome was sparse and limited and, thus—to properly characterize the patterns of LD and the haplotype variation across the whole genome—the International HapMap project was launched in 2001[20].

The HapMap project[21, 22] sequenced 269 individuals in 4 populations (Europe, Africa, China, and Japan), identified around 6 million new genetic variants, and mapped the recombination frequency, LD pattern, and haplotype variation across the genome, together forming the haplotype map of the human genome. They demonstrated that the recombination often happens within small sections of the genome, known as recombination hotspots. The LD in between the recombination hotspots is generally high (see the example in Figure 3). Consequently, the haplotype variation within the blocks could be captured with high confidence using only a few genetic variants, known as tagging variants, Indeed, the project estimated, that in European or Asian populations, only around 500,000 tagging variants are needed to capture most of the common variation with R^2>0.8; in African populations, the number of tagging variants was estimated as being slightly over 1 million[22]. These observations, together with the pioneering open-data sharing policy, facilitated the design of cost-effective genotyping chips, thereby enabling cheap genotyping and modern genome-wide association analyses.

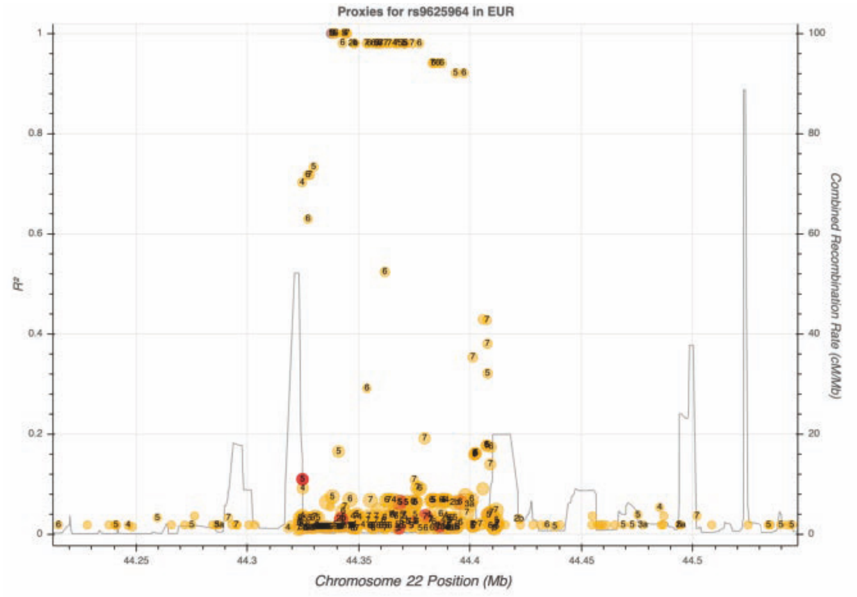


Figure 3. Example of an LD pattern at region 44.2-44.5 Mb in chromosome 22 in the European (EUR) population. The dots show the correlation coefficient (R^2, y-axis on the left) between the variant rs9625964 and other variants located in different positions at x-axis. The color of the dots shows whether the variant is coding (red) or non-coding (yellow). The grey line shows the combined recombination rate (y-axis on the right). The example demonstrates that there are 7 variants in complete LD and several other variants in high LD with the chosen variant forming a horizontal pattern (haplotype block) that is bounded by regions with a high recombination rate. The figure is produced with a public webtool LDproxy (ldlink.nci.nih.gov)[23].

2.2 POPULATION GENETICS AND GENETIC STRUCTURE

Population genetics studies the genetic variation within and between populations, including its the evolutionary effects. Understanding population genetics provides tools and solutions for multiple fields conducting genetic analyses. First, population genetics helps in answering the fundamental questions about our origin, and provides complementary tools for multidisciplinary studies of human history. Second, population genetics supports studies of complex diseases to control for genetic structure. Third, population genetics is utilized in forensics for identification of individuals and their relationships. This section familiarizes the reader with the field's basic terms and focuses on the core concept of this thesis, genetic structure.

2.2.1 POPULATION GENETIC PROCESSES AND TERMS

The basic principles of inheritance and population genetics were established already in the 19th and 20th centuries, well before the direct access to DNA variation. Below, the information of those processes and terms is based on the book *Principles of Population Genetics*[24] and lecture material *Population and Quantitative Genetics*[25], if not otherwise indicated.

Mutation is the process that physically alters DNA sequence producing genetic variants. Therefore, it is ultimately mutation that increases genetic variation both within and between populations. However, as mentioned earlier, the rate of mutation in humans is relatively low, and thus mutation affects genetic variation relatively slowly, on a long time scale. The mutations that happen in a germline are inherited through generations and work as the fundamental material for evolution.

Genetic drift is the term used to describe the random fluctuation of allele frequencies between generations due to chance only. In diploid organisms, genetic drift arises from the random sampling of alleles in meiosis leading to a varying number of alleles to be transmitted to the next generation, compared to the previous generation. The magnitude of genetic drift depends on the population size and allele frequency: the smaller the population, the larger the changes in relative allele frequencies may be. A simple example with three populations of different sizes, N = 10, 100, and 1000, shows that if the allele count increases by 5 between generations t and t+1, the changes in relative allele frequencies are 0.25, 0.025, and 0.0025, respectively. If no other evolutionary processes are involved, genetic drift eventually leads to a fixation of one allele and to the loss of the other alleles and genetic variation. The expected time required for a biallelic locus to become fixed is proportional to the minor allele frequency.

As opposed to genetic drift**, natural selection** is the evolutionary force causing non-random change in the genetic composition of a population. Natural selection arises from the non-random imbalance in the reproductive success of genetically varying individuals. Natural selection works through phenotypic variation increasing the frequency of genetic variants associated with favorable phenotypic variation (positive selection) and decreasing the frequency of deleteriously affecting variants (negative selection). Thus, the effect of natural selection on genetic variation varies depending on the effect of variation on phenotype.

Migration, i.e., the movement of individuals to other living habitats, is a process that introduces new genetic variation to the habitat by either mixing with the existing population or by establishing a new population. The effect of migration on genetic variation is again proportional to the relative sizes of each of the migrating and the original groups, and may be effective on a short time scale.

As explained above, changes in population size affect the level of genetic variation by providing the above processes an opportunity to act. Especially, the rapid changes in population size have a major effect. **Bottleneck effect** is used to describe the decrease in genetic variation due to the extreme reduction in population size. The genetic variation is decreased because the remaining variation is usually chosen by random leading to the loss of especially rare frequency variation at the same time as genetic drift amplifies the possibility of fixation of alleles in small populations. **Founder effect** is a special case of a bottleneck effect where a small founding group establishes a new population leading to genetic consequences similar to a bottleneck effect. In contrast, exponential **population growth** increases genetic variation as it provides more opportunities for both existing and novel variants to be transmitted to the next generation. The effect of exponential population growth on genetic variation has been observed in humans as an excess of rare variants compared to the model with a constant population size[26].

2.2.2 GENETIC STRUCTURE

Genetic variation in a population is not always homogeneously distributed but can show heterogeneous patterns known as genetic structure (Figure 4). Genetic structure arises from non-random mating. In most natural populations, individuals mate with others who are geographically close to them resulting in patterns where geographically closely located individuals are also genetically more similar. In turn, geographically distant groups become genetically more distant, even in the absence of physical barriers, and this phenomenon is called isolation by distance. The effect of geographic distance on genetic variation and structure is further shaped by the evolutionary processes described above. Another term, tightly connected with genetic structure, is admixture that describes genetic mixing between ancestral groups and can be used as a concept to characterize genetic structure.

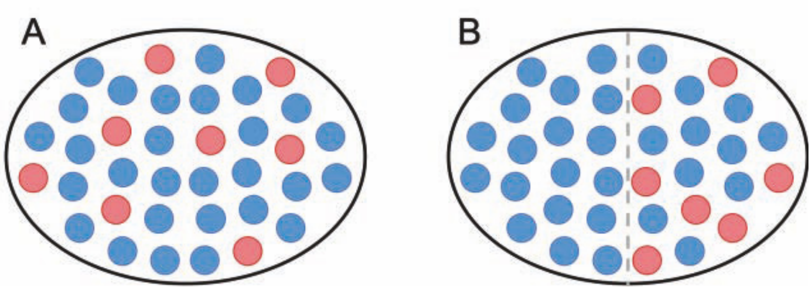


Figure 4. An example of a homogeneous (A) and a heterogeneous (B) population. Circles represent individuals in a population and the color represent which genetic variant, blue or red, they carry. While the red variants are randomly distributed in the population A, in population B, they are spatially clustered in the right-hand side of the population. Therefore, the population B can be described as having dichotomous genetic structure illustrated by the grey dotted line.

In human populations, multiple studies with varying methods have identified clear patterns of genetic structure and its correlation with geographic distance on a world-wide scale[27-31] as well as in more detail within Europe[32-37], Africa[38-43], Asia[44-46], North America[47-49], and South America[50-52]. Additionally, the methodological advancements and increase in sample sizes have enabled identification of extremely detailed genetic structure within single countries, such as in Great Britain and Ireland[53-56], Japan[57-59], France[60], Spain[61] and the Netherlands[62] among others. Although geographic distance remains the main factor behind genetic structure, other factors, such as culture, language[63], and religion[64, 65] can play a role in the formation of genetic structure as well. The effect of cultural factors on genetic variation are often complex and interconnected[66]. For example, the increase in the frequency of lactase persistence, caused by a genetic variant in *MCM6* gene, has been shown to be tightly connected to the spread of dairy farming in Europe[67].

Population genetic structure changes over time, as evolutionary processes and demographic events shape the genetic composition of groups. Therefore, by understanding the changes in genetic structure we can make interpretations about population history. Changes in genetic structure and population history are examined either by directly comparing time series of ancient DNA samples[68-70] or by comparing only modern populations to each other. When using only modern populations, the idea is to look for traces of past demographic events from modern genomes, such as bottleneck effect and genetic drift altering allele frequencies[40]. Such studies and approaches have portrayed ancient and historic events hundreds and thousands of years ago[62, 68, 71, 72]. Historic events and consequent changes in genetic structure of more recent history, for example within the last 100 years, have, however, been less studied.

2.2.3 GENETIC ANCESTRY

In spoken language, ancestry refers to any preceding origin of an individual or a group, and is often loosely used to connect ancestors, relatives, and different characteristics via both genetic and cultural inheritance. However, in genetics, ancestry is strictly defined as the biological origin delivering genetic information. Mathieson and Scally[73] have recently elaborated the terminology further by explicitly defining genealogical ancestry and genetic ancestry separately. Genealogical ancestry of an individual encompasses all those ancestors that are connected to the individual via a family tree, and thus the theoretical number of the genealogical ancestors is $2^{G,}$ where G is the number of generations back. Instead, the genetic ancestry refers only to those genealogical ancestors from whom the individual has directly inherited genetic material. Because of the recombination and random transmission of haplotypes, at particular locus, the probability of inheriting genetic material from the genealogical ancestors halves in every generation, and consequently the number of genetic ancestors (being roughly $2 \cdot (22 + 33 \cdot (G - 1))$) quickly becomes many fewer than the number of genealogical ancestors[74]. Identification of genetic ancestors is challenging in practice, and thus it is often approximated by estimating admixture and genetic similarity to some existing reference groups. In this thesis, the genetic ancestry is estimated via an individual's genetic similarity to predefined reference groups.

2.2.4 METHODS TO STUDY GENETIC STRUCTURE

Numerous methods to evaluate and visualize different aspects of genetic population structure exist (see the reviews[75-77]). The methods can be broadly, although not unambiguously, categorized based on the type of information they utilize. Most classical methods, such as F_{ST}-measure, principal component analysis (PCA) and STRUCTURE program[78, 79], were developed before large-scale genotype data and adequate computational resources were common, and thus they are traditionally based on the allele frequency differences of independent genetic variants (although multiple extensions to manage large sample sizes and LD have been implemented). In turn, the modern haplotype-based methods utilize information from rare and tightly linked variants. They have also been shown to gain more power to detect fine-scale genetic structure than the allele frequency-based methods[4, 80]. Because rare and tightly linked variants unfold more information about the recent evolutionary processes, the haplotype methods are also more suitable for studies of recent demographic events than the frequency-based methods[81]. In the following, I familiarize the reader in more detail with two classical methods, F_{ST}-measure and PCA, and one haplotype-based method, chromosome-painting, that are used in this thesis.

F_{ST}-measure

The most frequently utilized statistic for population genetic distance is the F_{ST}-measure. Sewall Wright developed a set of F-statistics, i.e., fixation indices, to measure genetic variation within and between populations at the turn of the 1950s[82] and later these statistics have been extensively utilized and extended[83]. The main idea of the F_{ST}-measure is to compare genetic variation within a subpopulation to the total genetic variation of the whole population as

$$F_{ST} = 1 - \frac{2p_s(1-p_s)}{2p_T(1-p_T)}$$

where p_S is the allele frequency of the subpopulation and p_T is the allele frequency in the whole population[25]. Moreover, F_{ST} can be extended across multiple loci and populations by simply averaging sub- and total population variances before calculating the ratio. F_{ST} is often reported for two populations, in which case it measures pairwise-F_{ST} that is the difference in allele frequencies between the populations relative to the total variation. F_{ST} between most human populations have been observed to be small (<0.1)[84] which, in practice, means that the genetic variation between these population is tiny compared to all variation among humans.

Principal component analysis

Another cornerstone for population genetic studies is principal component analysis (PCA). PCA is a statistical technique to extract major axes of variance, known as principal components (PCs), from multidimensional data. These components describe a decreasing amount of variance, are orthogonal to each other, and thus are convenient for data visualization. PCA was introduced to genetics first by Cavalli-Sforza and colleagues in the 1970s[32] and again for modern whole-genome data by Patterson and colleagues in the 2000s[85]. In genetics, PCs can be perceived to describe orientations in which the individuals show the most genetic variation. In statistical terms, PCA is applied on a $N \times S$ data matrix where N is the number of samples, S is the number of independent genetic variants and the matrix entries, g_{il}, are the genotypes for sample i in locus l. The standard protocol to perform PCA[25, 85] on biallelic data is to first standardize each entry of the data matrix by extracting the mean genotype $2p_l$ and scaling with the corresponding standard deviation as

$$M_{il} = \frac{g_{il} - 2p_l}{\sqrt{2p_l(1-p_l)}}.$$

Second, the standardized data matrix M is transformed into a sample covariance matrix X as

$$X = \frac{1}{S}MM'.$$

Now the major axes of sample variance are found by conducting an eigenvalue decomposition on the covariance matrix X. In practice, PCA is performed with existing software such as EIGENSOFT[85, 86] and Plink[87] and the PCs are visualized in pairs on two-dimensional scatter plots. If PCs encompass genetic structure, it can be detected as non-random patterns, often triangular or U-shaped, on a PCA plot. However, the interpretation of the pattern is not straightforward as sample size, possible correlation between variants, and the strength of evolutionary processes have varying effects on PCA. Naturally, multiple extensions for standard PCA have been implemented including, for example, fast implementation for big data[88] and an implementation to account for correlation between samples[89] among others.

Chromosome-painting, FineSTRUCTURE, and SOURCEFIND

In 2012, Lawson and colleagues introduced a haplotype-based method, chromosome painting[4] to identify fine-scale populations from genetic data. Later, it was successfully used to describe the fine-scale genetic structure of the British Isles[53]. The method is a collection of software tools that build on a matrix that summarizes the haplotype-based relationships of individuals in the data. This coancestry matrix is created by the ChromoPainter[4] program and it can be further utilized to identify population structure in a PCA-like manner or by clustering individuals into discrete populations with the FineSTRUCTURE[4] program. Furthermore, additional programs utilizing the coancestry matrix to estimate admixture and ancestry, such as GLOBETROTTER[90] and SOURCEFIND[52], have been published. In what follows, the general idea of the programs utilized in this thesis, ChromoPainter, FineSTRUCTURE and SOURCEFIND, are introduced.

Chromosome painting aims to identify the number (and the length) of shared haplotypes between the individuals in the data and this information is assumed to capture rich information about the underlying genealogies (Figure 5A). The theoretical framework for chromosome painting is based on an algorithm of Li and Stephens[91] which models haplotype chunks as a Hidden Markov Process. Broadly, the method compares one individual (recipient) at a time to all other individuals (donors) and estimates, for each locus, which (one or more) of the donor individuals are the closest to the recipient individual on that position (Figure 5B-F). By modeling all loci sequentially using the Hidden Markov Model, chromosome painting produces local estimates of haplotype chunks and with whom each chunk is most recently shared. These chunks are then summed across the genome generating the coancestry matrix where the rows describe how many chunks each recipient individual shares with each donor individual.

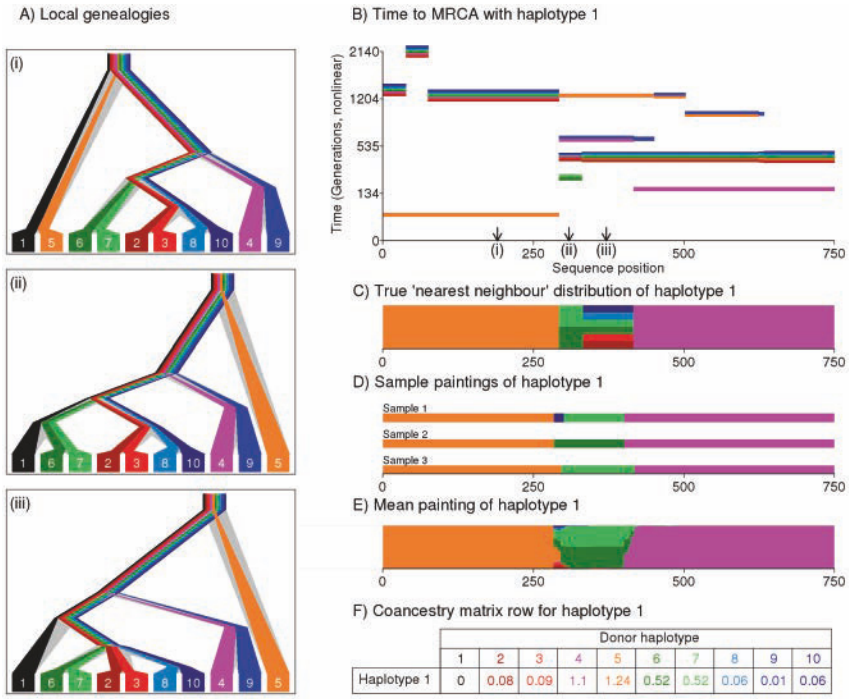

A) Local genealogies
(i)
1 5 6 7 2 3 8 10 4 9
(ii)
1 6 7 2 3 8 10 4 9 5
(iii)
1 6 7 2 3 8 10 4 9 5
B) Time to MRCA with haplotype 1
Time (Generations, nonlinear)
2140
1204
535
134
0
0 (i) 250 (ii) (iii) 500 750
Sequence position
C) True 'nearest neighbour' distribution of haplotype 1
0 250 500 750
D) Sample paintings of haplotype 1
Sample 1
Sample 2
Sample 3
0 250 500 750
E) Mean painting of haplotype 1
0 250 500 750
F) Coancestry matrix row for haplotype 1
Donor haplotype
1 2 3 4 5 6 7 8 9 10
Haplotype 1 0 0.08 0.09 1.1 1.24 0.52 0.52 0.06 0.01 0.06

Figure 5. A schematic example of a chromosome painting method where the haplotype of individual number 1 (recipient) is compared to nine other individuals (donors). Different colors represent the different individuals. A) Recipient individual has different underlying genealogies for three loci (i, ii, iii), orange individual being the closest in locus i; dark and light green in locus ii; greens, reds and blues in locus iii. B) The time to the most recent common ancestor (MRCA) as a function of sequence position matching the genealogies and representing haplotype patterns. C) The true distribution of the closest haplotype chunks that is being estimated. D) The example chromosome paintings produced by ChromoPainter algorithm. E) The mean chromosome painting averaged over sample paintings. F) The output row in the coancestry matrix, where the number of haplotype chunks are summarized over the genome, demonstrating that the recipient individual shares the most chunks with the orange individual. Reprinted from PLOS Genetics 8(1), Lawson *et al.* (2012) *Inference of population structure using dense haplotype data* under the Creative Commons Attribution License.

The genetic structure of the coancestry matrix can be identified either by performing a PCA on it (explained briefly in Section 4.4 or in detail in [4]), or by clustering the individuals into fine-scale populations with FineSTRUCTURE[4]. FineSTRUCTURE implements, first, a Markov chain Monte Carlo (MCMC) method to assign individuals into discrete populations, and second, builds a hierarchical tree to represent the relationships of the populations. In general, MCMC tries to find the optimal population assignment by altering the current assignment either by splitting or merging existing populations, or moving individuals between the populations, and then accepting the new assignment with a probability based on a likelihood ratio. If the assignment is accepted, the next new assignment is produced based on the accepted one, otherwise it is produced based on the current assignment. The end result is a sequence of different population assignments. If the MCMC algorithm is continued long enough, the assignments are expected to converge around the optimal assignment given the data. The assignment that produces the largest posterior probability is then used as the final population clustering and is used in the tree-building phase. The FineSTRUCTURE tree is built “bottom up” in such a way that the genetically closest fine-scale populations are successively merged into larger populations. While the FineSTRUCTURE tree does not represent any true genealogy or evolutionary model, it has been shown to successfully capture the general genetic structure at multiple levels[4].

The results of ChromoPainter and FineSTRUCTURE can be further utilized to infer admixture proportions as implemented in the SOURCEFIND program[52]. The idea of SOURCEFIND is to compare the chromosome painting of the test individual, i.e., the row of the coancestry matrix, with the chromosome painting of the predefined reference groups to infer admixture proportions. To simplify the calculation process, the chromosome paintings are first summarized into copying vectors which sums the fractions of the genome copied from the donor individuals belonging to the same genetically homogeneous groups defined, e.g., by FineSTRUCTURE. The copying vector of the test individual is then modelled as a weighted mixture of the reference groups where the weights are inferred as the admixture proportions. To find the optimal weights, SOURCEFIND uses an MCMC-algorithm that has been shown to converge to the optimal admixture proportions, given the data[52].

2.3 GENETIC STRUCTURE IN THE GENETIC STUDIES OF COMPLEX DISEASES

2.3.1 GENOME-WIDE ASSOCIATION STUDY

The main approach for determining genetic factors underlying complex diseases is a genome-wide association study (GWAS). GWAS seeks to identify genetic variants underlying a disease (or a trait) by performing a statistical test separately for each variant through the whole genome. More specifically, GWAS fits either a linear or logistic regression model on the data, and estimates an effect size, i.e., the strength of the correlation, and a p-value, i.e., the strength of the statistical significance, between the variant and the disease. If the p-value of the variant is under the genome-wide significance level, (typically $< 5 \cdot 10^{-8}$), the variant is considered to be associated with the disease. Essentially, GWAS identifies a list of variants that show statistical association that most often is only tagging the effect of a real causal variant. During the GWAS era, almost 180,000 associations for a wide range of diseases and traits have been identified (the GWAS catalog[92], March 2020). While the GWAS associations alone are not a proof for causality, there are many examples where a GWAS has been able to pinpoint biological pathways relevant for pathological mechanisms of a disease or to identify promising therapeutic targets (examples summarized in[93, 94]).

As both allele frequencies and the incidence of complex diseases are affected by multiple factors, GWAS are sensitive for confounding. Confounding is a statistical term to describe a setting where a possibly unknown, third factor affects both the outcome variable (here the disease or trait) and the predictor (here variant), which causes a spurious association between the outcome and the predictor, and can lead to a false interpretation about causality. In GWAS, age, sex, and technical factors, such as batch and genotyping plate effects, are routinely controlled for. In addition, because both the allele frequencies and diseases can show geographic variation, it was clear, already at the arrival of the first GWAS studies, that the genetic structure can cause serious confounding in genetic association studies and should be controlled for[95-97]. An exaggerated but classic example[98] can be given by imagining a GWAS on the ability to eat with chopsticks within a sample including individuals from both European and Asian background. Without a control for genetic structure, this study would tag multiple loci associated with Asian ancestry, not because they would be biologically relevant but because the Asian culture and genetic ancestry were correlated. A standard method for controlling genetic structure in GWAS is a PCA-based correction[86] that adjusts the regression model with, e.g., the top 10 to 20 PCs. Additionally, multiple other methods have been developed[76] and the modern methods utilize linear mixed models assessing directly the genetic relationship of the samples[99]. These standard methods have been successful in identifying significant associations that have been replicated in other cohorts[100] while under- or over-correction can still exist (Figure 6).

The current association studies are also facing some further challenges related to genetic structure. These challenges include three main issues. First, most GWAS are being conducted in populations with European ancestry which limits genetic diversity and the opportunities for new biologically relevant discoveries and translation to other populations[2, 3, 101]. Second, GWAS are increasing in size as the large-scale studies with 500,000 samples and more are being collected, for example, in the UK (UK Biobank)[102], Japan (Biobank Japan)[103], USA (The Million Veterans Program)[104] and in Finland (www.finngen.fi). Consequently, studies analyze rarer variants that show more subtle and localized genetic structure[105, 106], and thus more sophisticated methods for controlling genetic structure are needed. Third, the GWAS results from non-genome-wide-significant variants are also used in subsequent analyses, including genomic prediction (section 2.3.2), requiring new standards especially for effect estimates. Together, these factors demonstrate that the proper understanding of genetic structure and diversity is essential also in the future.

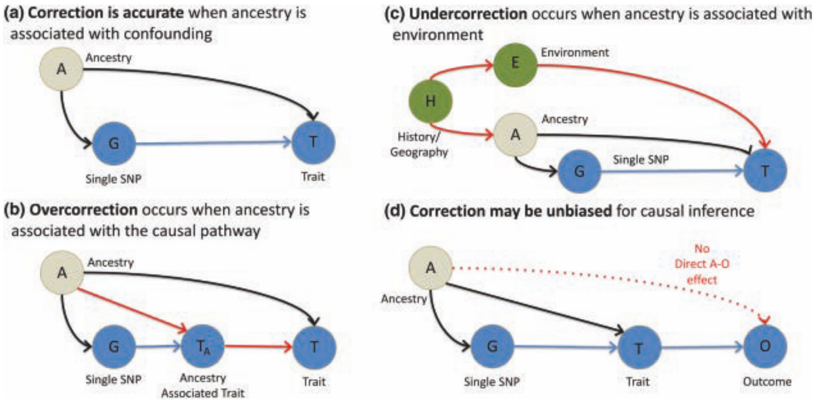

(a) Correction is accurate when ancestry is associated with confounding
A
Ancestry
G
T
Single SNP
Trait
(b) Overcorrection occurs when ancestry is associated with the causal pathway
A
Ancestry
G
T_A
T
Single SNP
Ancestry Associated Trait
Trait
(c) Undercorrection occurs when ancestry is associated with environment
E
Environment
H
History/ Geography
A
Ancestry
G
Single SNP
T
(d) Correction may be unbiased for causal inference
A
Ancestry
No Direct A-O effect
G
T
O
Single SNP
Trait
Outcome

Figure 6. Models of correction with genetic ancestry for variant-trait associations. Reprinted from Human Genetics 139, Lawson *et al.* (2020)[107] *Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?* under the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

2.3.2 POLYGENIC SCORES

The results of GWAS are being increasingly utilized in genomic prediction to estimate and evaluate the genetic risk of an individual in developing a specific disease. Currently, genomic prediction is implemented via polygenic scores (PSs) that summarize the genetic risk estimate of multiple variants into a single measure. Because the effects of single variants on complex diseases are small, a large number of variants are included in the PS to gain a good risk prediction. For example, the early PSs of coronary artery disease (CAD) utilized information from 13[108] and 28[109] genetic variants, but showed only modest, if any, improvements in risk prediction when combined with traditional risk factors. Later, a PS with 49,310 variants improved risk prediction by around 5% among those over 60 years old[110]. Now, the modern PSs employ LD-information to summarize millions of variants across the whole genome[111] and utilize, e.g., 6,412,950 variants for CAD[112]. These highly polygenic scores have demonstrated, for instance, that individuals in the top 8% of PS distribution match the risk of carriers of a known monogenic mutation in coronary artery disease[113] and that the PSs improve the risk prediction, especially for the early-onset of the disease, compared to the traditional risk factors alone[112] supporting future utilization of PSs in clinical setting and disease prevention. Similar results have also been obtained for other diseases, such as for type 2 diabetes[112, 113] and breast cancer[114].

While PSs have been utilized in multiple ways, there remain two main challenges, tightly linked to genetic structure, that hinder their utilization in health care and in between populations comparisons. First, because a PS is calculated as a sum of an individual's genotypes weighted by the effect estimates from a GWAS over a large number of variants, even a tiny but consistent directional bias in effect estimates, harmless for a single variant, can accumulate a substantial bias for PSs. As was discussed above, such a bias can easily arise from poorly controlled genetic structure. As an example, the GWAS meta-analysis of adult height by the GIANT consortium [115] was shown to include severe biases[116, 117] that had already led to an apparently false conclusion about a strong differentiation in height in Europe[118]. In addition, a study utilizing UK Biobank has reported that latent, fine-scale genetic structure is present in GWAS results, even after adjusting for 40 PCs biasing PSs[119]. Similarly, fine-scale structure is found to affect PSs of Biobank Japan[59] demonstrating that the need for controlling subtle genetic structure in GWAS still exists. Recent results have shown that detecting genetic structure with haplotype-based methods is a promising approach to control for such biases in GWAS[62].

Second, the prediction accuracy of PSs decreases the more genetically distant the target population is from the original GWAS population[2, 3]. This phenomenon is not only a result of poorly adjusted population structure but also a result of other GWAS characteristics. As GWAS have more power to find common as opposed to rare variants, the associated variants are skewed towards the variants common in the original GWAS population. For example, there is a concerning observation that the overall GWAS results are more polymorphic in European populations compared to other populations than what is expected based on the known overall genetic variation[3]. This imbalance is due to the dominance of European ancestry among the GWAS samples. Further, environmental factors and LD-patterns between populations might vary, causing uncertainty to the effect estimates of tagging variants. As a result, comparison of PSs between populations is challenging and has been shown to lead to unrealistically large or even contradictory differences between distant continental populations[3, 120] but there is limited information about how similar problems manifest within the populations with fine-scale genetic structure. Additionally, as the prediction of PS depends on the ancestry, the interpretation of PS for admixed individuals is challenging and methods employing local genetic ancestry have been proposed[121]. Altogether, these challenges demonstrate that we need to understand fine-scale genetic structure not only to be able to control for it but also to understand the limitations of the spurious relationship between the target and the GWAS samples before translation of PS in clinical use.

2.4 POPULATION OF FINLAND

The Finnish population has been actively utilized in the studies of human genetics for decades. The following sections summarize the history and other characteristics of the population outlining the reasons for this exceptional interest from the international research community.

2.4.1 HISTORY OF FINLAND

Information on the population history of Finland is limited to the time after the last Ice Age, which ended approximately 11,000 years ago, as no reliable archaeological discoveries exist before that. The following times are often divided into prehistorical periods, where only archaeological evidence exist, and into historical times when also written documents are available. Figure 7 gives a broad overview of the main periods and events during both the prehistorical and historical eras of Finland. If not otherwise stated, the information on Finnish prehistory is based on the book *Muinaisuutemme jäljet: Suomen esi- ja varhaishistoria kivikaudelta keskiajalle*[122] whereas the historical period is based on the book *Suomen Historia Jääkaudesta Euroopan unioniin*[123].

As the southern and western parts of Finland were covered by water after the Ice Age, the oldest human remains have been found in the southeastern and northern parts the country, and have been dated between 10,000 and 11,000 years old. The southeastern remains are believed to originate from people who arrived from the South and East, whereas the northern remains are believed to originate from the people who arrived along the coast of Norway. Over the millenia since then, it is presumed that people have continuously inhabited Finland[124]. However, the density of inhabitation has varied and a wide range of different cultural, and possibly genetic, influxes have occurred. The Stone Age populations were based on the hunter-gatherer lifestyle and new cultures were adopted along with small migrating groups. A large number of archaeological findings from 7,000 to 5,000 before the present (BP) have demonstrated a wide spread of the Comb Ceramic and related cultures across the whole country. For example, the Typical Comb Ceramic culture covered large geographic regions from the southwest corner all the way up to the Bothnian Bay and further to the East (Figure 8A)[122, 125]. Outside of Finland, the Comb Ceramic culture was typical for Northern and Eastern Europe around the Baltic Sea, suggesting that Finland had close connections to these neighboring areas. This period was followed by a gradual adoption of agriculture and dairy farming at the end of the Stone Age. Agriculture and dairy farming have, furthermore, been associated with the Corded Ware culture (5,000–4,000 BP) and with the Kiukainen culture (4,500–3,500 BP). But unlike the Comb Ceramic Cultures, these other cultures have been discovered mainly from the southwestern corner of Finland (Figure 8B)[122, 125]. Abroad, the Comb Ceramic related cultures are found to be widely distributed around the Baltic Sea, Central Europe, and Russia[126].

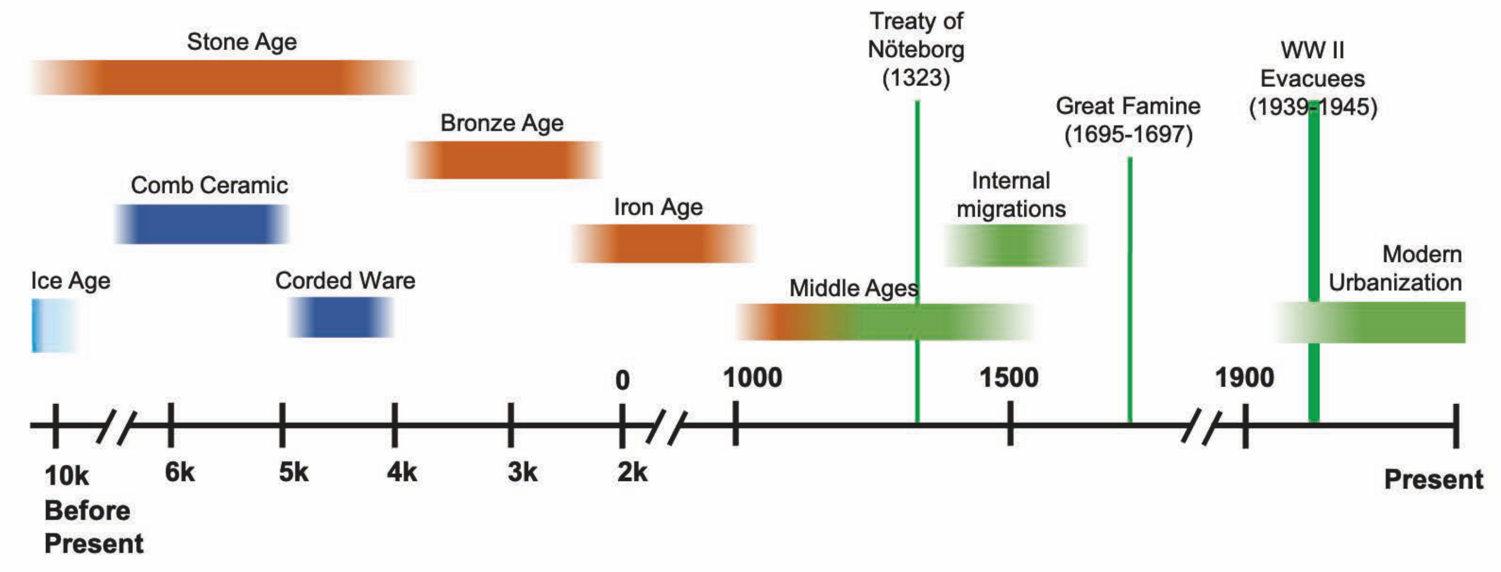


Figure 7. Timeline for the main archeological periods and historical events in Finland. The Ice Age is marked with the color cyan, the archaeological periods with the color brown, ancient cultures with the color blue, and historical events with the color green color. The numbers below the timeline correspond to the years before the present, and numbers above the line correspond to calendar years in the common era.

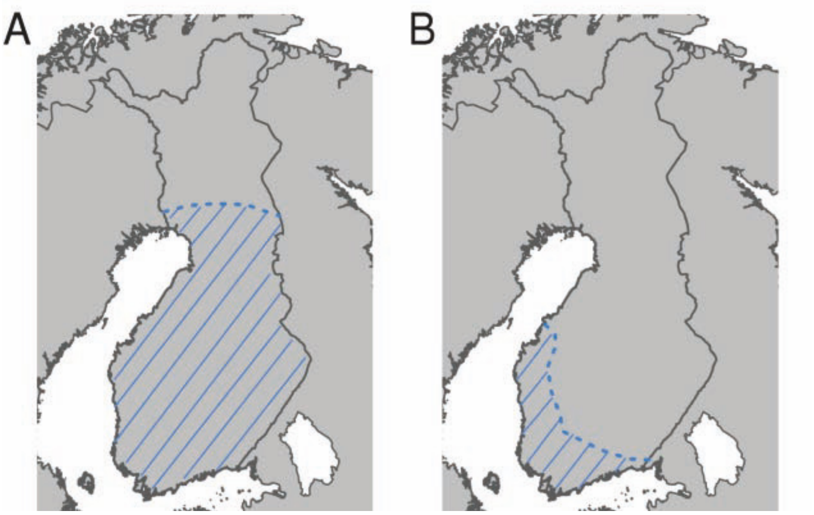


Figure 8. The approximate spread of A) the Typical Comb ceramic culture (5,900–5,500 BP) and B) the Corded Ware culture (5,000–4,000 BP) and the Kiukainen culture (4,500–3,500 BP) in Finland during the Stone Age according to the archaeological findings[122].

The Bronze Age in Finland (4,000–2,500 BP) was characterized by a reduction in the number of archaeological findings as well as by regional differences: the coastal region had connections to southern Scandinavia while the inland region had trading connections to the East all the way to the Volga region and the Urals. In turn, during the Iron Age (2,500–1,000 BP), the population size started to increase due to stabilized agricultural practice, especially in the southwestern parts of Finland. The concluding centuries of the Iron Age were outlined by the Viking movements and crusades, which both were operated from Scandinavia and once again influenced especially the southwestern parts of Finland.

The Middle Ages represents the shift from prehistorical to historical times in Finland and was characterized by the power struggle between the eastern and western realms, Novgorod and Sweden. With the Treaty of Nöteborg in 1323, struggle relaxed between the two powers; the western parts of the country became subject to taxation for Sweden and the eastern parts for Novgorod. The treaty border of 1323 was Finland's first known eastern border, but it was partly only loosely defined and followed approximately from the southeastern corner to the western coastal region (Figure 9B). It is also noteworthy that a considerable migration of Swedish people to the coastal regions of Finland occurred in the 13th and in the beginning of the 14th century. The permanent settlement was scattered around the country, but the southern and western coastal regions up to the Bothnian Bay showed denser and more continuous settlement (Figure 9A). These regions are often referred to as the early settlement region in the genetics literature, but it is noteworthy that the eastern and northern parts of the country, often known as the late settlement region, were not uninhabited either. The Middle Ages could also be highlighted for the increase in population size, which was estimated to be around 300,000 at the time. The increasing population size in turn increased the pressure to establish new farming sites and the settlement became more continuous. People from Savonia are often highlighted as being the most eager to move to new areas during the 15th and 16th centuries.

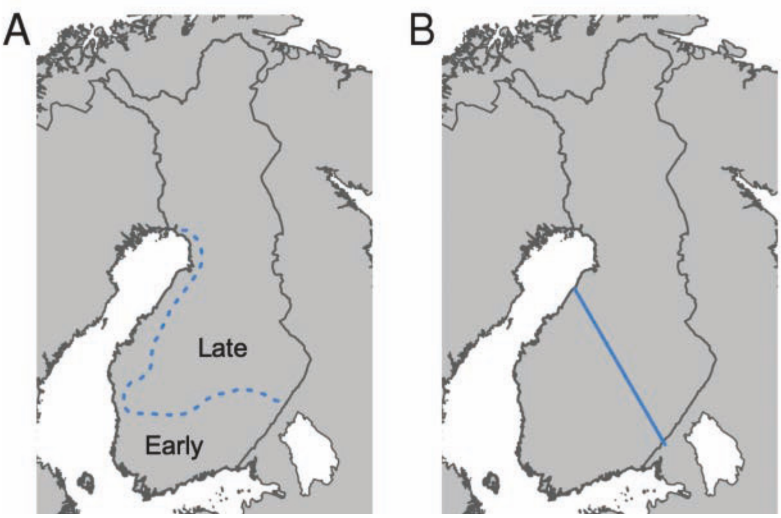


Figure 9. A) The permanent settlement in Finland during the Middle Ages concentrated earlier on the southern and coastal regions of Finland and continuously covered the whole of Finland only later. B) The border of the Treaty of the Nöteborg in 1323 followed approximately from the Vyborg Castle to the coast of Ostrobothnia, but its exact location is not clearly known.

The centuries following the Middle Ages were harsh for the Finnish population. Several wars between Sweden and the reunited Russian Empire were conducted on Finnish territory and the Great Famine killed up to one third of the population between 1695 and 1697. Consequently, the population size did not noticeably increase until the latter part of the 18th century. Starting from the end of the 18th century, the population expansion triggered further movement within the country. Again, especially the Savonian people moved to North Ostrobothnia and to the

Oulu region on Finland's western coastline along the Bay of Bothnia, but there was movement also to the South and to the existing cities.

The 20th century remarked the rise of nationalism and the independence in Finland. After centuries of being ruled by either Sweden or the Russian Empire, Finland declared its independence in 1917. While the transition to independence was itself peaceful, the following years were characterized first by the Finnish Civil War (1918) and later by the three episodes of World War II (WWII) in Finland: The Winter War (1939–1940), the Continuation War (1941–1944) and the Lapland War (1944–1945). The war during the 1940s induced the largest known internal migration within the country as large areas in the eastern and northern parts of Finland were ceded to the Soviet Union. In consequence, around 400,000 individuals (around 11% of the population) were relocated to other parts of the country. The relocations were controlled and well documented by the Finnish authorities, and while the plan was to distribute the evacuees evenly across the country, the southern and western parts of the country gained around 70% of the evacuees[127]. However, the evacuees are known to have moved further after the initial relocations. For example, a large number of the evacuees moved away from the Southern Ostrobothnia region right after the war while Southern Finland gained additional evacuees[127]. The population history of Finland from the end of the 20th century, as well as at the beginning of the 21st century, has been mainly characterized by urbanization. In Finland, urbanization first increased the number of small, local towns in the 1950–2000, and only fairly recently, has this urbanization centralized into a few major cities[128].

2.4.2 LANGUAGE IN FINLAND

Finland has two official languages: Finnish with 4.8 million speakers (87%) and Swedish with 300,000 speakers (5%). In addition, 7% of the population speak languages such as Sami, Russian, or Estonian, among others[129]. In short, the Finnish language belongs to the Uralic language family and, more specifically to the Finno-Ugric subgroup together with, for example, Estonian and Hungarian[130]. The Finnish language shows dialectal variation that is well documented in the Finnish Dialect Atlas[131]. The dialects are classically divided into the Eastern and Western dialects, and further into seven or eight subdialects: Southwestern (and Southwest transitional), Tavastian, Southern Ostrobothnian, Mid- and

North Ostrobothnian, Far-Northern, Savonian, and Southeastern[132, 133]. More recently, this classification has been confirmed and further elaborated by applying quantitative tools to linguistic data[134, 135].

2.4.3 GENETICS IN FINLAND

The work in this thesis has focused on a major genetic group within Finland that is naively referred to as Finns here. However, Finland, and the Finns, also include other groups that show varying levels of genetic relatedness to the major group studied here. For example, the Sami people are a native Finno-Ugric-speaking group that has inhabited the northern parts of Finland, Sweden, Norway, and the northern Kola Peninsula in Russia for millennia and show distinct genetic characteristics from the major group[136-138]. Then again, the Swedish-speaking Finns show a close relatedness to the Finnish-speaking Finns and can be genetically identified only when samples from the Swedish population are included in the analysis[139, 140]. Furthermore, the Finnish gene pool is getting more diverse in consequence of the world-wide immigration. For example, in 2019, 8% of the Finnish population had a foreign background[141]. Of this percentage, the largest groups with over 20,000 individuals had a background from Russia, Estonia, Iraq, or Somalia.

Finnish population in global context

The early population genetic studies of European populations placed Finns as an outlier population on the genetic map of Europe[142]. Later, PCA-based studies[34-36, 143, 144] have elaborated the position of Finns at the northeastern edge of the continuum of European genetics background, rather than as an outlier. The studies on a world-wide scale have highlighted the same: Finns are close to the European superpopulation but show a strong eastern affinity[6, 145]. For example, Finns show more Siberian ancestry than Estonians[146]. Although, depending largely on the populations analyzed, the genetically closest populations to Finns are the geographic neighbors: Estonians[144, 146], Karelians in Russia[146], and Swedes[34-36, 144].

Genetic structure within Finland

Genetic structure within Finland has been rigorously studied and its main feature, the division into East and West (more specifically into Northeast and Southwest), was first characterized by the blood group studies of

Harri Nevanlinna and colleagues in the 1970s. The studies described frequency differences of blood group markers not only between East and West, but also between detailed municipality-level regions[147, 148]. These observations were followed by more direct analyses of genetic markers, as haplogroup studies of Y chromosome demonstrated a strong East–West divergence[137, 149-155]. In turn, the first haplogroup analyses of mitochondrial DNA did not show strong population differences or a deviation from European diversity[156, 157], but with more data, subtle regional differences in mitochondrial DNA have also been observed[154, 155, 158, 159]. Consistently, the autosomal data have repeated the above results of the main genetic structure as being between East and West[139, 140, 144, 160], and has shown, for example, that the genetic distance, measured with pairwise-F_{ST}, is larger between Eastern and Western Finland than between Germany and Great Britain[140]. However, these previous studies have evaluated the genetic differences of East and West either by comparing geographically distinct samples from the opposite corners of Finland and dismissing the central parts of Finland, or by comparing geographically defined regions, such as provinces. Thus, there is no detailed understanding of where the genetic borderline between East and West exactly lies, and in turn how admixed people near the borderline are.

Beyond the East-West division, there is also evidence of further genetic substructure in Finland. However, the information on this topic is more limited. Many studies utilizing samples around Finland performed analyses with only a few genetic loci, such as haplogroups in mitochondrial DNA or Y-chromosome, and have reported subtle frequency differences between Finnish provinces[151, 153, 154, 158, 161, 162]. Moreover, the autosomal scans with tens of thousands of unlinked variants have suggested considerable fine-scale genetic structure. In the study of Salmela *et al*. (2008)[140], a PCA-based analysis separated part of the samples from different provinces both in Eastern and Western Finland, and the study of Jakkula *et al*. (2008)[160] showed striking substructure within Northern Finland, especially between the different parts of Lapland. More recently, and together with the studies presented here, haplotype-based methods have allowed a more detailed examination of the genetic structure within Finland[81, 163]. As the amount of data continues to increase, it is expected that we will discover further details about the genetic structure and admixture in the future[75].

Ancient DNA studies

The advancements in DNA technologies have allowed the analysis of DNA from ancient and historical samples, and during the last ten years, the studies of ancient DNA (aDNA) have provided unique opportunities to understand human history. The studies of aDNA have, for example, elaborated the mutation rates and Neanderthal admixture in modern humans[164]. In Europe, one of the oldest remains of modern humans found from Romania (dated 37,000-41,000 BP) surprisingly did not show genetic proximity to modern Europeans[165]. Instead, the modern Europeans are currently seen as a mixture of three ancient groups, contributing to the gene pool at different time points after the Ice Age[68, 72]. The base of the modern European populations is built on groups of hunter-gatherers who widely inhabited Europe soon after the Ice Age. This gene pool was admixed with the early farmers from the Near East (~8,500 BP) and with Yamna (a.k.a. Yamnaya) ancestry from the Eurasian Steppe (~3,000–7,000 BP). However, the contributions and details of this simplistic model vary across Europe and, especially in Northeastern Europe, there is substantial evidence of additional contributions from the East and from Siberia[71, 146, 166-168].

Unfortunately, the studies of ancient populations in Finland are almost nonexistent. The soil in Finland does not preserve human remains well[169] and thus there is a limited number of samples suitable for aDNA analyses. The first major study including Finnish aDNA examined 7 samples from Levänluhta burial site, and showed that these samples, dated between 300 and 700 CE, exhibit closer affinity to the modern Sami people than modern Finns[167], providing little insight into the birth of the modern genetic structure. In turn, a larger study with 70 ancient mitochondrial-DNA (mtDNA) samples, from three different ancient and historical periods, detected a considerable spatial and temporal heterogeneity in mt-haplogroups: The change of the main mt-haplogroups of Finland, U and H, was observed such that, during the Iron Age and the medieval era in Finland, haplogroup U was the most common group in the southwestern corner of Finland. But among the modern samples, H is the most common in the southwestern parts—and this change in U/H ratio was almost 6-fold[159]. In the eastern parts of Finland, an opposite change was observed, although less dramatic, suggesting that especially the early population history of Finland is not well understood. Presumably, the additional aDNA (including mtDNA, Y-chromosomal, and genome-wide) data combined with the modern samples will elaborate the population history of Finland in the future.

Finns in the genetic studies of diseases

In the 1970s, researchers observed that Finland exhibit multiple rare diseases that are, in practice, absent from the other countries of the world. These diseases were soon observed to be inherited and were defined as the Finnish Disease Heritage, a group of rare, genetic diseases almost exclusive to Finland[170, 171] (http://www.findis.org/). The quest to understand these diseases was one of the initializing forces in studying genetics in Finland. The same mechanisms that produced the Finnish Disease Heritage have more broadly led to a phenomenon where some rare variants have been enriched in the population (along with the loss of many other variants)[172] and are therefore easier to identify in a GWAS in the Finnish population than in other populations. Consequently, and because of the active research community, Finns have been included in several international consortia studying the genetics of disease and traits, including for example GWAS meta-analyses of coronary artery disease[173], schizophrenia[174] and adult height[115]. Currently, a large-scale biobank collection of 500,000 Finnish samples is being carried out and their genotyping in the FinnGen Project (www.finngen.fi) highlights the continuation of genetic studies in Finland.

2.4.4 GEOGRAPHIC VARIATION OF DISEASES AND TRAITS IN FINLAND

Despite the top-level health-care system in Finland[175], the Finnish population shows geographic variation in the disease prevalence and incidence rates. In addition to the geographic clustering of many rare diseases of the Finnish Diseases Heritage[170, 176], also many complex diseases show geographic variation. As an example, the general morbidity index of the Finnish institute for health and welfare (summarizing information on coronary artery disease, cerebrovascular diseases, cancers, musculoskeletal diseases, dementia, mental-health problems, and accidental injuries) is considerably higher in Eastern than in Western Finland[177]. Coronary artery disease (CAD) is one of the main drivers of these regional differences (together with musculoskeletal diseases and mental-health problems) showing 1.6 times higher incidence rates in Eastern than in Western Finland between 2013 and 2015[178]. Indeed, the dramatic difference in CAD incidence rates between

Eastern and Western Finland initiated, already in 1972, a community-based program called the North Karelia Project aiming to prevent cardiovascular disease, and was later expanded to cover other regions of Finland forming the foundation for the National FINRISK Study[179] utilized in this work (see section 4.1). Together the projects successfully managed to decrease coronary mortality by over 80% between 1972 and 2014[180]. Yet, after over 40 years of favorable changes in lifestyle and improvements in health care, there still exist some regional differences in CAD incidence rates[178]. However, not all existing geographic differences are between East and West in Finland. For example, ulcerative colitis has been reported to be more common in northern than in the southern parts of Finland[181]. While multiple traditional factors, such as socioeconomic status and lifestyle, are known to play a key role in the risk for complex disease, the role of population genetic differences in regional health discrepancies is not well understood.

3 AIMS OF THE STUDY

The Finnish population is one of the most studied populations in human genetics. Due to the ongoing large-scale biobanking efforts, advancements in polygenic risk prediction and attempts to integrate genetic information as part of health-care systems, it is increasingly important to thoroughly understand the genetic structure in Finland. In this thesis, the main goal is to characterize how the genetic structure develops geographically and how it is connected to the complex diseases and traits in Finland. More specifically, this thesis aims to

1. Characterize the fine-scale genetic structure (Study I),
2. Track the changes in the genetic structure throughout the 20th century (Study II), and
3. Map the geographic variation and population structure-related bias in polygenic scores (Study III)

in Finland.

4 MATERIALS AND METHODS

4.1 STUDY SUBJECTS

This thesis utilizes samples from the National FINRISK Study (hereafter FINRISK). FINRISK is a survey of the Finnish adult population (24- to 75-year-olds) to examine chronic and noncommunicable diseases and their risk factors. The study includes a comprehensive questionnaire and health examination that capture a wide range of information from measured blood pressure to self-reported lifestyle factors, such as sleep and physical exercise. In addition, FINRISK has collected biological samples including DNA.[182]

FINRISK consists of surveys that have been collected at 5-year intervals, beginning since 1972. In this thesis, Studies I and III utilize data from the survey of 1997 and Study II uses the surveys of 1992, 1997, 2002, 2007, and 2012. While the FINRISK collection has been centralized into five collection regions, shown in Figure 10, the birth places of the study participants readily cover most parts of the country. However, there are considerably more samples collected in North Karelia (NKA) and North Savonia (NSA) compared to other parts of the country. Table 1 presents the number of samples in each of the study regions, for each study separately. As FINIRSK was the largest and the most comprehensive Finnish population cohort including genetic information from all parts of Finland at the time this work was started, it was the natural choice for studying genetic structure in Finland.

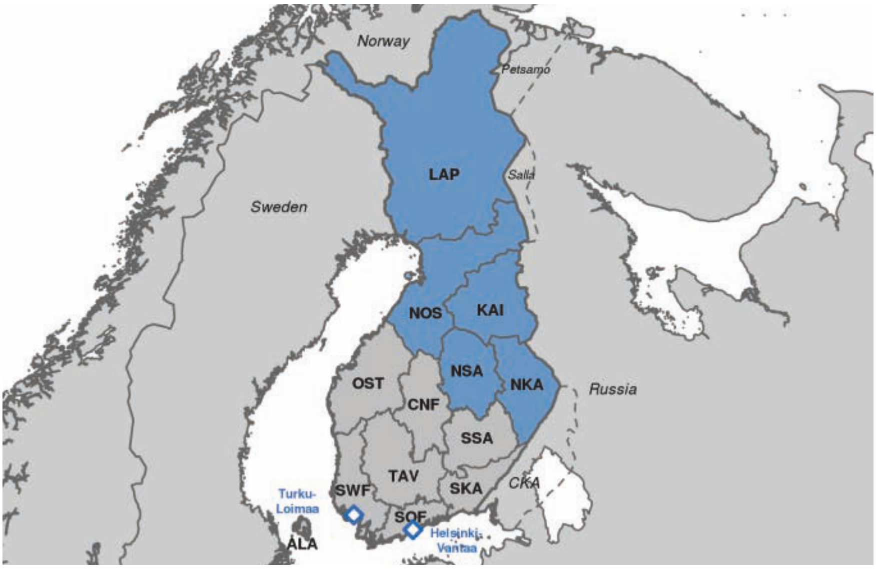


Figure 10. The sample collection regions of the National FINRISK Study (blue) and the study regions examined in this thesis (in bold). Abbreviations correspond to those presented in Table 1.

Table 1. Number of samples per region in Studies I, II and III.

Region		**Study I**	**Study II**	**Study III**
LAP	Lapland	38	1,010	38
NOS	North Ostrobothnia	206	2,291	382
KAI	Kainuu	57	726	140
NSA	North Savonia	139	3,060	592
NKA	North Karelia	139	3,088	587
CNF	Central Finland	45	420	45
SSA	South Savonia	69	622	90
SKA	South Karelia	47	442	49
OST	Ostrobothnia	84	555	85
TAV	Tavastia	71	833	75
SWF	Southwestern Finland	109	3,073	226
SOF	Southern Finland	38	1,870	67
ÅLA	Åland	0	8	0
CKA	Ceded Karelia	0	465	0
Total		1,042	18,463	2,376

4.2 QUALITY CONTROL

I utilized FINRISK samples that had been genotyped with Illumina HumanCoreExome genotyping chip that consists of over 500,000 rare and common genetic variants and were called with zCall[183] preceding this study. The usage of only one type of genotyping chip assisted in merging the data but does not insulate them from batch effects or genotyping errors. To ensure the high quality of the data, I performed the following variant and sample related quality-control steps. For each variant, I calculated minor allele frequency (MAF), Hardy-Weinberg equilibrium (HWE) p-value (that quantifies whether the number of observed genotypes matches with the number expected based on the allele frequency), and the proportion of missing samples. I excluded the variants, if their MAF was under 5% or HWE p-value was under 10^{-6} or missingness was over 1%. For each sample, I calculated the proportion of missing variants and the rate of heterozygosity. Samples were excluded if they showed large deviations from other samples in these measures. In addition, I excluded one individual from each pair of related individuals (3rd degree) and individuals who themselves, or whose parents were born outside of Finland, or who did not have sufficient location information. Quality measures were calculated using PLINK1[87] and PLINK2[184]. In Study II, relatedness was calculated using KING[15]. For each study, the specific description of the quality-control steps and thresholds are given separately in the original publications.

4.3 GEOGRAPHICALLY UNIFORM SAMPLE SETS

For Studies I and II, I selected a geographically uniform subset of the samples to ensure a robust identification of fine-scale populations. In Study I, the uniform sample set of 1,042 individuals were selected by placing a grid of 25 km on Finland and by randomly sampling a maximum of 5 individuals from each grid square. In Study II, the set of 2,741 samples were selected in a two-step process that sequentially excluded individuals with the largest number of geographically closest neighbors. On the first round, individuals were excluded until no individual had more than 15 neighbors within 5 km proximity reducing the number of samples in large cities. On the second round, the number of individuals were further reduced in such a way that each individual had at maximum 40 individuals within 30 km radius, resulting in a geographically uniform sample set. The exclusions were performed in R[185].

4.4 PRINCIPAL COMPONENT ANALYSIS

I performed principal component analyses (PCA) on our data to both identify possible quality issues and to study the genetic structure of our samples. In Study I, PCA was run using SmartPCA program of the EIGENSOFT library[85]. In Study II, PCA was run in PLINK1.9[87]. PCAs were performed with a set of LD-independent variants ($R^2 < 0.2$): 61,598 variants in Study I and 56,661 in Study II. In Study II, 31 samples were excluded as they showed closer affinity to the word-wide populations than to the Finnish samples in PCA.

For the method comparison in Study I, I performed a custom PCA on the haplotype-based coancestry matrix as described in Lawson *et al.* (2012)[4]: coancestry matrix was first modified by adding the column sums to the diagonal, subtracting the column means from all the elements and last multiplied by its transpose. This resulted in a normalized, symmetric matrix on which PCA was performed with the eigen-function in R[185].

4.5 PAIRWISE-F_{ST}

Pairwise-FST measures were calculated using SmartPCA program of the EIGENSOFT library[85] using command 'fstonly' on the same set of independent variants that were used for PCA.

4.6 HAPLOTYPE-BASED ANALYSES

While principal component analysis is a useful tool for detecting genetic structure, it excludes information on tens of thousands of variants that are correlated via genetic linkage. To include information from these linked variants I utilized a haplotype-based method called chromosome painting. Here, I used chromosome painting implemented in ChromoPainter and its companion program, FineSTRUCTURE, to identify fine-scale populations.

4.6.1 GENOTYPE PHASING

Before identifying shared haplotypes, ChromoPainter requires that the samples are jointly phased. Phasing is a computational process to determine which variants were inherited together in one chromosome and thus it determines individual haplotypes. I phased the data with SHAPEIT2[186] using an average European effective population size (11,418) and recombination map from the HapMap phase II[22].

4.6.2 CHROMOPAINTER: CHROMOSOME PAINTING

As briefly introduced in section 2.2.4, chromosome painting identifies the number and the length of shared haplotypes between the recipient and donor individuals. In study I, I used ChromoPainter[4] version 0.0.4 to paint all 1,042 samples against all other samples. First, I estimated parameter values for global switch and mutation rates utilizing chromosomes 1, 9, 15, and 22, and following Leslie *et al.* (2015)[53]. The individuals were then painted using the estimated parameters. Other parameters were kept at the default values. In Study II, I used ChromoPainter version 2.0 similarly as before, but all 18,463 individuals were painted against a subset of 2,741 geographically evenly distributed individuals. ChromoPainter outputs a coancestry matrix, which is a non-symmetric genetic relationship matrix, and is further used to identify fine-scale populations and estimate genetic ancestry.

4.6.3 FINESTRUCTURE: IDENTIFICATION OF FINE-SCALE POPULATIONS

To identify fine-scale genetic populations, I utilized a clustering approach implemented in the FineSTRUCTURE[4] program. FineSTRUCTURE is designed to first identify small, genetically homogeneous groups based on ChromoPainter's coancestry matrix, and second build a bifurcating tree by merging these groups into higher-level populations.

In Study I, I ran FineSTRUCTURE version 0.0.5 on the coancestry matrix of 1,042 individuals. To identify fine-scale populations, I ran FineSTRUCTURE's MCMC algorithm using 1,000,000 burn-in-iterations, 1,000,000 sample iterations recorded every 10,000th iteration, first by allowing only two populations, and second allowing any number of populations. In Study II, I ran ChromoPainter on 2,741 individuals similarly to that above, except that I used version 2.0.1 and 2,000,000 sample iterations. Additionally, before tree-building, I re-assigned a part of the individuals into new populations to improve the overall posterior probability by following the procedure of Lawson *et al.* (2012)[4]. The FineSTRUCTURE-tree was built using the default options.

4.6.4 BUILDING THE TVD-TREE

While FineSTRUCTURE merges small genetic groups into a reasonable tree-structure, it is affected by the sample size. To build an alternative tree that utilizes the haplotype information but is not affected by the sample size, I employed the measure of total variation distance (TVD) described in Leslie *et al.* (2015)[53]. TVD compares two copying vectors, *a* and *b*, i.e., the rows of coancestry matrix, to each other as

$$TVD_{a,b} = 0.5 \cdot \sum_{i=1}^{K} |a_i - b_i|,$$

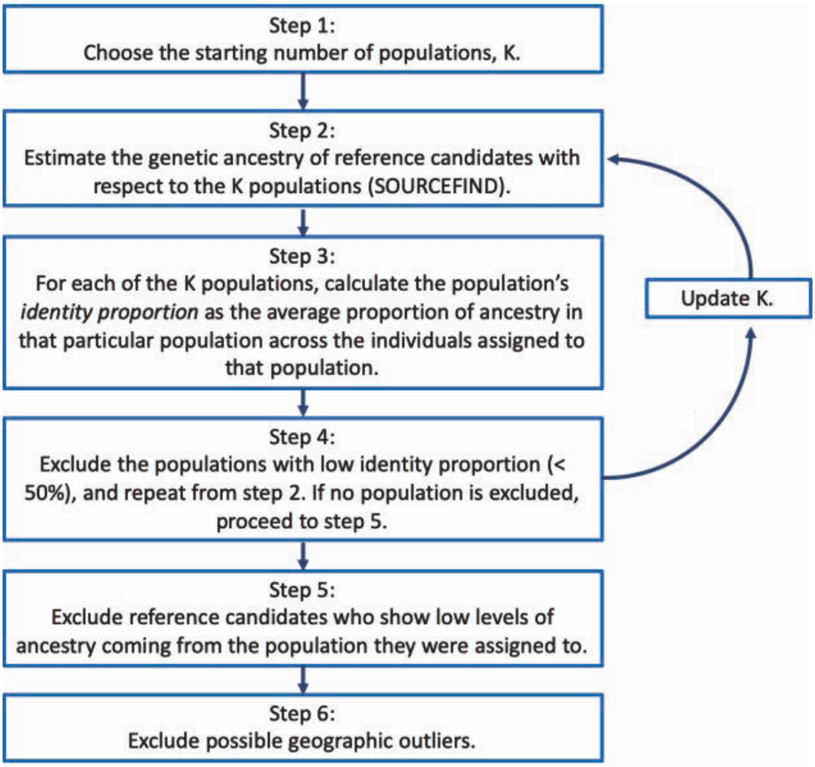
where a_i (b_i) is the copying proportion from a refence population *i*. Here, I used the 17 fine-scale populations defined by FineSTRUCTURE-tree as the number of reference populations (K=17). To build a tree based on the TVD measure, I started with the 17 fine-scale populations, calculated the pairwise TVDs, and recursively merged the two populations that showed the smallest TVD.

4.7 ANCESTRY ESTIMATION

4.7.1 IDENTIFYING REFERENCE GROUPS

Ideal reference groups for genetic ancestry estimation would be ancestral, genetically homogeneous (i.e., do not exhibit substructure), genetically independent (i.e. are not admixed) and would each represent one well-defined group such as a geographic region. To identify reference groups that mimic the above characteristics, we developed a workflow illustrated in Figure 11. The procedure utilizes predefined populations. First, it iteratively excludes admixed populations (steps 2–4), and, second, excludes individuals that do not unambiguously represent the population to which they were assigned (step 5), and, finally, excludes the individuals if they are single, extreme geographic outliers of their own population (step 6). Note that an intensive exclusion of geographic outliers should be avoided and manually checked, so as not to produce reference groups based on prior expectations.

To identify the reference groups for ancestry estimation in Study II, I started with 2,741 geographically evenly distributed individuals whose parents were born within 80 km from each other. I identified the fine-scale populations among these 2,741 reference candidates using FineSTRUCTURE (see section 4.6.3) and focused first on the two main populations, East and West. I estimated that the identity proportions (see Figure 11, step 3) of both East and West populations were over 80%, and thus I continued to estimate each individual's own ancestry. I excluded 1,266 individuals whose both western and eastern ancestry proportions were under 95% and 3 individuals who were geographic outliers. These reference groups are illustrated in Figure 16A (in Results section 5.2.1) and are referred to as reference groups in reference set 2 (hereafter abbreviated as refset).


Step 1:
Choose the starting number of populations, K.

Step 2:
Estimate the genetic ancestry of reference candidates with respect to the K populations (SOURCEFIND).

Step 3:
For each of the K populations, calculate the population's *identity proportion* as the average proportion of ancestry in that particular population across the individuals assigned to that population.

Step 4:
Exclude the populations with low identity proportion (< 50%), and repeat from step 2. If no population is excluded, proceed to step 5.

Step 5:
Exclude reference candidates who show low levels of ancestry coming from the population they were assigned to.

Step 6:
Exclude possible geographic outliers.

Figure 11. Workflow to identify robust reference groups for ancestry estimation with fine-scale populations.

Second, I started with 15 fine-scale populations and estimated their identity proportions. Out of the 15 populations, 5 showed low identity proportions (under 50%) and were excluded. The ancestry estimation was repeated with the remaining populations, and this time the remaining populations showed either around 70% or around 80% identity proportions. I continued with all 10 reference populations, and, in addition, generated reference groups with the 6 populations that showed around 80% identity proportions. For both sets of reference groups (i.e., refset 10 and refset 6), I excluded heterogeneous individuals showing under 70% of ancestry coming from the population they were assigned to. Additionally, I manually excluded the individuals who were located outside the core region of their reference group. The reference groups in refsets 6 and 10 are shown in Figure 16 (in Results section 5.2.1) panel B and C.

4.7.2 SOURCEFIND: ESTIMATION OF ANCESTRY PROPORTIONS

Genetic ancestry was estimated using SOURCEFIND[52] v2. In our approach, I ran SOURCEFIND for all 18,463 individuals that were painted against the 2,741 donor individuals. For MCMC iterations, I used 50,000 burn-in iterations followed by 150,000 sample iterations and recorded every 5,000th sample. The ancestry proportions were then estimated as an average over the 30 recorded samples. SOURCEFIND was run separately for refsets 2, 6, and 10. The genetic ancestry was separately estimated for the reference individuals such that the reference individual itself was excluded from the reference group while other parameters were kept as described above.

4.7.3 SIMULATING INDIVIDUAL ANCESTRIES

To test the usability of our reference groups I performed simulations with real-world data. First, I identified ancestor candidates that matched the geographic location of the reference groups and then simulated offspring by modeling recombination between the ancestor chromosomes.

The ancestor candidates were identified among the FINRISK samples that were not part of the 2,741 reference candidates but whose parents were still born within 80 km from each other. For the first set of simulations between East and West, I identified ancestor candidates whose parents were born in SWF or OST (A-West ancestors) or in CKA, NKA, KAI, NKA or LAP (A-East ancestors). For the second set of simulations, I identified ancestors from 7 more detailed regions, SWF, OST, LAP, NKA, KAI, Kuusamo/Pudasjärvi and CKA. In addition, the ancestor candidates were assessed as having homogeneous genetic background based on PCA. Figure 12 shows the geographic locations of the ancestor candidates.

I simulated offspring by randomly sampling ancestors among the candidates defined above. Each ancestor candidate was used only once. I simulated two kinds of offspring: those who had all their *N* ancestors coming from one ancestor group and those whose 1 ancestor came from one population and the other *N-1*, where N = $2^{\#generations}$, ancestors came from another population. This allowed us to estimate the proportion of detectable ancestry as the function of generations back in time. In detail, I simulated the haplotypes of the offspring by sampling recombination events between the ancestor haplotypes and randomly sampling one of the novel haplotypes to be transmitted to the offspring. To sample the recombination events, I used the recombination map from the HapMap phase II.

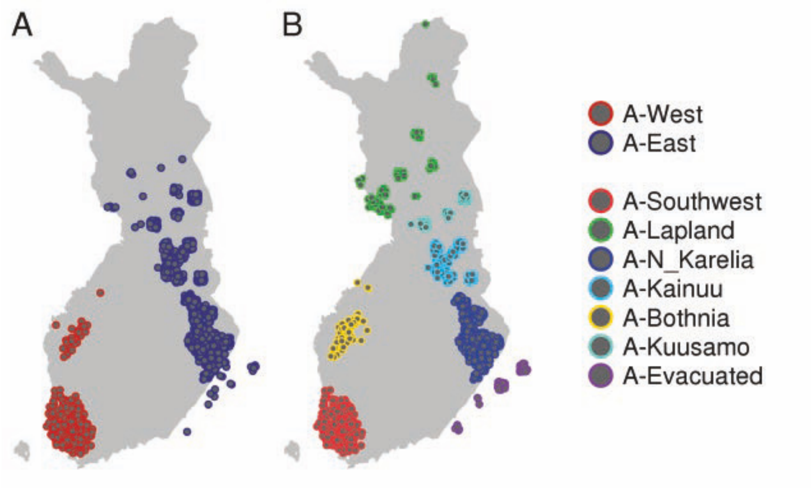


Figure 12. Location of ancestor candidates used for simulating offspring.

4.8 POLYGENIC SCORES

4.8.1 SUMMARY STATISTICS AND VARIANT FILTERING

To study the geographic distribution of polygenic scores and its connection to the genetic structure in Finland, we targeted 5 complex diseases and 3 quantitative traits. For each disease or trait, we used summary statistics from a large, international GWAS meta-analysis as listed in Table 2.

Summary statistics were filtered for minor allele frequency (<0.01), p-value (>0.05), imputation quality (<0.9), number of samples or cohorts in meta-analysis (<0.9 of the maximum possible), and by removing the region of the major histocompatibility complex. The multi-allelic variants were also excluded. The remaining variants were filtered to include LD-independent variants using an LD-clumping method with 500 kb window and 0.1 threshold for R^2 using PLINK 1.9.

Table 2. Characteristics of summary statistics used to build polygenic scores (PSs). The column 'GWAS N' shows the number of samples (cases/control) in the corresponding genome-wide association study.

Trait		**Study**	**GWAS N**	**Finnish samples**	**Variants in PS**
CAD	Coronary artery disease	CARDIoGRAM plusC4D[173]	60,801 / 123,504	5,825/ 5,639	19,597
RA	Rheumatoid arthritis	Okada *et al.* 2014[187]	18,136 / 49,724	--	32,736
CD	Crohn's disease	IIBDGC[188]	5,956 / 14,927	--	21,771
UC	Ulcerative colitis	IIBDGC[188]	6,968 / 20,464	--	23,513
SCZ	Schizophrenia	PGC[174]	36,989 / 113,075	--	30,311
WHR	Waist-hip ratio	GIANT[189]	224,459	~16,000	13,727
		FINRISK	24,919	24,919	43,252
BMI	Body mass index	GIANT[190]	322,154	~23,000	12,742
		UKBB[191]	337,199	--	75,979
		FINRISK	24,919	24,919	44,920
HG	Height	GIANT[115]	253,288	~23,000	27,066
		UKBB[191]	337,199	--	113,079
		FINRISK	24,919	24,919	50,536

4.8.2 POLYGENIC SCORES AND GENETIC RISK MAPS

Polygenic scores were calculated for each FINRISK individual *i* as a sum of genotype, x_{ij}, weighted by the effect estimate, $\hat{\beta}_j$, of a variant *j* as

$$\mathrm{PS_i} = \sum_{j=1}^{M} x_{ij}\hat{\beta}_j.$$

PSs were later scaled by the mean and standard deviation of 1,042 geographically evenly distributed individuals from Study I.

Genetic risk maps were generated by overlying a grid on top of the map of Finland and for each map point, *p*, I calculated a weighted averaged over all individual PS as

$$\mathrm{PS_p} = \frac{1}{\mathrm{r_{Tot}}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\mathrm{PS_i}}{\mathrm{r_{ip}^2}},$$

where r_{ip} is the distance between individual *i* and grid point *p*, and $\mathrm{r_{Tot}} = \sum_{\mathrm{i}} \frac{1}{\mathrm{r_{ip}^2}}$ is the sum of the weights. For each point, $\mathrm{PS_p}$ also included a pseudo individual with an average PS and a distance of 50 km to the point to deal with the uncertainty in regions with low sample sizes. Additionally, to avoid high variance in weights, I used a minimum value of 50 km in r_{ip}. The width of the grid was set to 10 kilometers. Genetic risk maps were plotted in R using map data from GADM data base https://gadm.org/.

4.8.3 POLYGENIC AND PHENOTYPIC DIFFERENCES

To quantify polygenic and phenotypic differences between East and West, we first defined the genetic groups using 2,376 individuals. The genetic East and West for this sample was defined using ChromoPainter and FineSTRUCTURE similarly to Study I defined above. The analysis assigned 1,604 individuals to East and 772 individuals to West.

The observed phenotypic difference was calculated for quantitative traits by adjusting the traits for sex, age, and age^2 using linear regression. WHR was additionally adjusted for BMI. The observed phenotypic difference was finally calculated as the difference in the mean of the residuals between the two groups in their original units, e.g., in cm for HG.

Polygenic score differences were calculated first in standard deviation units after scaling the PS with the subset of 1,042 (Study I) geographically evenly distributed samples. The phenotypic difference predicted by the PS was calculated for the quantitative traits only by first fitting the linear model where the phenotype was explained again by sex, age, and age^2. WHR was additionally adjusted for BMI. Second, the residuals from the first model were explained by the PS in a linear regression and the corresponding effect estimate for PS was used to transform the PS difference between East and West into phenotypic scale by multiplication.

4.8.4 DETECTION OF BIAS ACCUMULATION BETWEEN POPULATIONS

To detect bias accumulation between two populations in GWAS summary statistics, we developed an empirical approach that utilizes the least significant part of the variants in the summary statistics. The approach assumes that if there is a bias aligned with the population structure between the populations of interest, this bias is included in all variants equally, regardless of whether they associate with the phenotype or not. PS built on only weakly associated variants would accumulate the bias between the populations when more variants are included in PS but will not explain phenotype well.

I implemented the approach by filtering the summary statistics similarly to the original scores (see section 4.8.1), but included only the weakly associated variants with p-value above 0.5. Among these variants, I permuted the remaining p-values and performed the LD-clumping as above but set the p-value threshold to 1, so as not to exclude further variants. Then, I sampled randomly different numbers of variants (5,000; 10,000; 20,000; 40,000; 60,000 and 80,000) and calculated the corresponding PS. Each random PS setting was repeated ten times. The accumulation of these random PSs were compared to the expected results with truly zero effect estimates. For this, I generated an additional 1,000 simulated PSs, where their effect estimates were sampled from a normal distribution with a mean of zero and standard deviation corresponding to the standard error of the variant in GWAS. These 1,000 PSs were then used to define the 95%-confidence interval for the PS difference between the two populations.

5 RESULTS

5.1 GENETIC STRUCTURE IN FINLAND (STUDY I)

5.1.1 COMPARING STANDARD PCA AND CHROMOPAINTER

To motivate the usage of computationally intensive haplotype-based ChromoPainter over an efficient principal component analysis (PCA), we started with a comparison of these two methods. We ran a standard PCA on a set of unlinked variants and a custom PCA on a haplotype-based coancestry matrix of ChromoPainter program using 1,042 geographically evenly distributed FINRISK samples whose parents were born within 80 km from each other. Figure 13 shows that, while the standard PCA identifies West-East and South-North structure well, ChromoPainter separates, for example, individuals of Ostrobothnia (OST) from the individuals of Southwestern Finland (SWF) more clearly. We also compared the two methods quantitatively (see original publication of Study I) and showed that ChromoPainter clustered individuals from the northern and eastern parts of Finland more tightly together than the standard PCA, and only in Tavastia (TAV) did the standard PCA cluster individuals more tightly together than ChromoPainter. In the other regions, the difference was not significant. Taken together, these results demonstrated that haplotype-based methods show more potential in detecting fine-scale genetic structure than the standard PCA.

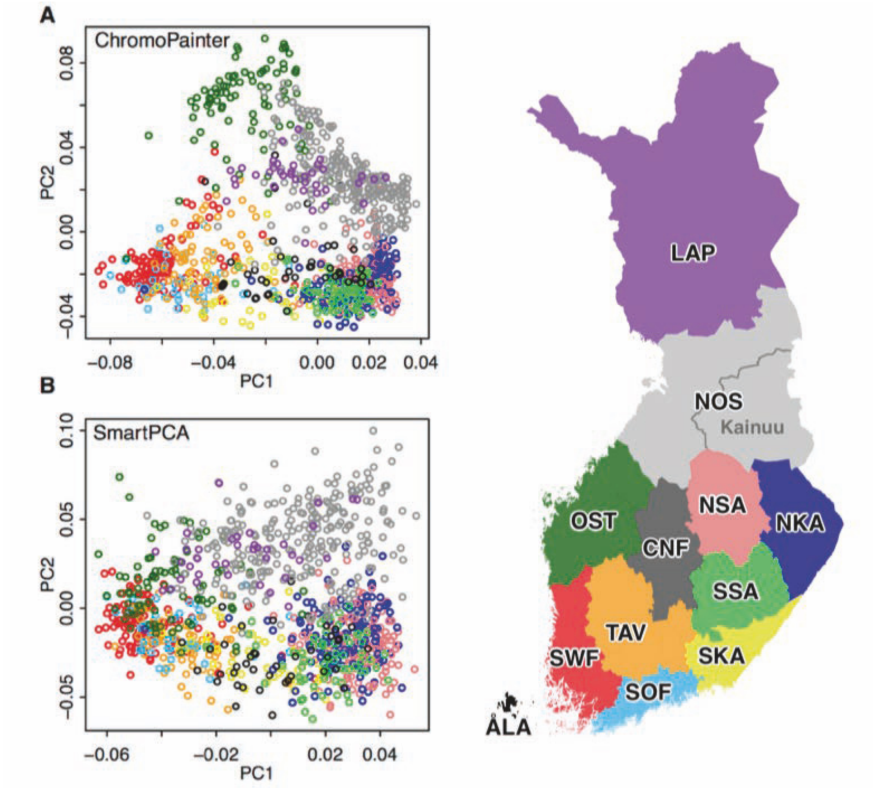


Figure 13. A) A custom principal component analysis (PCA) on a haplotype-based coancestry matrix and B) a standard PCA with unlinked variants for 1,042 Finnish individuals. The colors on the PCA-plots correspond to the individuals' geographic locations on the map. The figure is adapted from G3: Genes, Genomes, Genetics 7(10), Kerminen *et al.* (2017) *Fine-Scale Genetic Structure in Finland* under the Creative Commons Attribution License.

5.1.2 GENETIC BORDERLINE BETWEEN EAST AND WEST

We utilized haplotype-based methods and a geographically evenly distributed sample of 1,042 individuals to identify the main genetic division in Finland. The analysis identified two genetic populations, with a pairwise-F_{ST} of 0.002 (SE = $2 \cdot 10^{-5}$), one located in the western and the other in the eastern parts of the country (Figure 14). Because of the geographically dense and comprehensive sample, this division revealed, for the first time, the geographic location of the genetic borderline between the populations in East and West. The genetic borderline curves across Finland starting from South Karelia and ends at the coast of North Ostrobothnia.

We compared this genetic borderline to two historical borders: the early settlement border of the Middle Ages and the border of the Treaty of Nöteborg in 1323, as well as to the main dialect regions. The genetic borderline revealed substantial similarity to the approximated border of the Treaty of Nöteborg, while it had clear discrepancies with the settlement border in the central parts of the country. The dialect regions showed two distinct differences to the genetic borderline. In the South, the western population dominates the eastern dialect region, and near the city of Oulu, the eastern population dominates the western dialect region. In Study I, the genetic borderline was not compared to the spread of different cultures in the Stone Age, but by comparing the result in Figure 14 with the maps in Figure 8 (in section 2.4.1), it is notable that the genetic borderline also closely resembles the spread of the Corded Ware and the Kiukainen culture in the southwestern parts of the country.

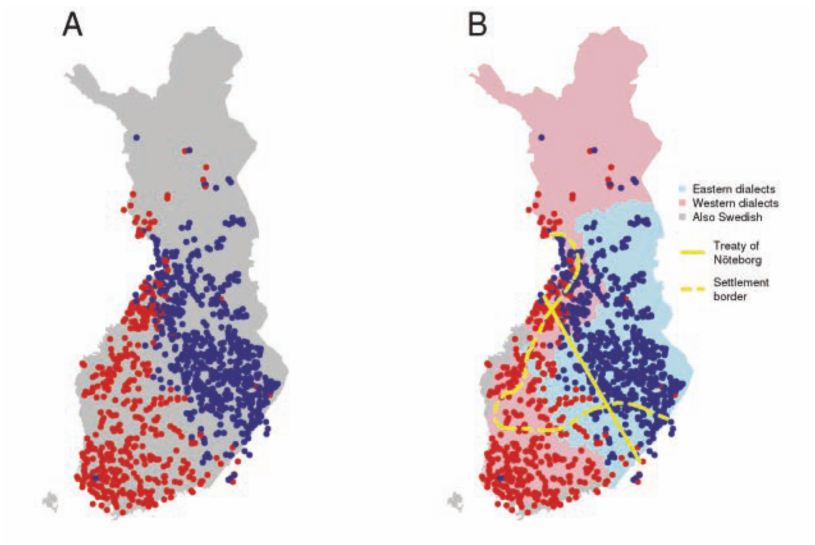


Figure 14. A) The main genetic populations in Finland are located in the East (blue) and West (red). B) The comparison of genetic populations to the borders of the early settlement region, Treaty of Nöteborg and to the main dialect regions. The figure is adapted from G3: Genes, Genomes, Genetics 7(10), Kerminen *et al.* (2017) *Fine-Scale Genetic Structure in Finland* under the Creative Commons Attribution License.

In addition to the analysis with 1,042 individuals (Study I), we repeated the analysis of the western and eastern populations in Studies II and III with two different sample sizes (N = 2,741 and 2,376, respectively). Despite the striking similarities between the three analyses, in the largest sample set, Lapland was clustered into the eastern population, while, in the two smaller analyses, it was part of the western population. Together with the results of PCA, where individuals from Lapland are distributed along the East-West component (Figure 13), it seems that the population of Lapland cannot be consistently assigned to either the western or eastern population as a whole; rather, the individuals form a continuum between West and East.

5.1.3 FINE-SCALE GENETIC POPULATIONS

To utilize the full potential of haplotype-based methods, we characterized the fine-scale genetic structure in Finland with an unprecedented level of detail. Using the sample of 1,042 individuals, the FineSTRUCTURE program identified 52 small, genetically homogeneous groups and merged them into higher-level populations. We verified that, at the level of 17 fine-scale populations, the populations were robust to varying sample sizes and had at least 25 individuals each.

The 17 fine-scale populations, shown and named in Figure 15, cover all other parts of Finland except Lapland and Åland Islands evenly, and are geographically tightly clustered with an exception of population P6. This population is mostly located in the southeastern corner of the country but is also scattered across the genetic borderline of East and West. When we compared the age distribution of the samples, we observed that the P6 population was slightly younger (median birth year 1957) than the other populations (median birth year 1950). In addition, P6 showed small and equal pairwise-F_{ST} values to all other populations independent of their status between East and West. Additionally, according to the ancestry analyses in Study II, the individuals in P6 were shown to share 70% of ancestry from the East and 30% from the West (unpublished results) suggesting that FineSTRUCTURE clustered together individuals of admixed background.

Because FineSTRUCTURE provides a tree-building algorithm that was observed to be affected by the sample size, we developed our own algorithm based on total variation distance (TVD) for inferring relationships between population and compared this newly generated TVD-tree to the original FineSTRUCTURE-tree. The TVD-tree showed more robust results with respect to the sample size than FineSTRUCTURE-tree in our data, and thus the relationships of the 17 populations are shown with the TVD-tree in Figure 15B. The TVD-tree shows that, after the division into East and West, both populations are divided in North-South direction, and the further splits demonstrate that the fine-scale populations are equally sized and geographically clustered around the country.

In addition to comparing the main dialect regions, we compared the fine-scale populations to the more detailed dialect regions (Figure 15C). This comparison showed that, in both the East and West, the genetic populations closely follow the dialect regions. The fine-scale populations most accurately follow the dialect border in South Ostrobothnia, Southwestern Finland, and in Lapland. In the Savonian dialect region, there are multiple genetic populations but they correspond well to the subdialect regions. On the contrary, the dialect region of Mid- and North Ostrobothnia contains individuals from multiple populations that also extend outside of the dialect region.

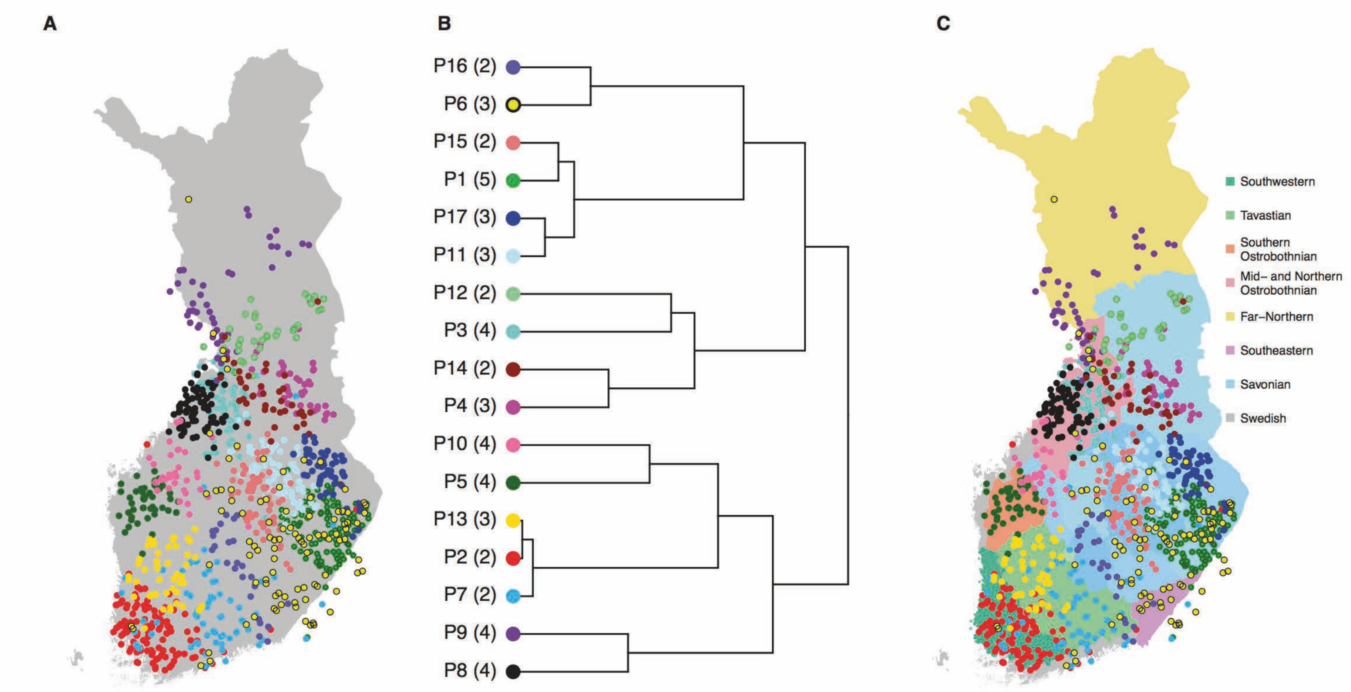


Figure 15. A) The 17 fine-scale genetic populations in Finland. B) TVD-tree presenting the relationships of the 17 populations. The letter P and a number (e.g., P6) represent the population label. The numbers in parenthesis demonstrate into how many subpopulations the population was further split in the complete FineSTRUCTURE-tree. C) Comparison of fine-scale populations to 7 subdialect regions. The figure is reprinted from G3: Genes, Genomes, Genetics 7(10), Kerminen *et al.* (2017) *Fine-Scale Genetic Structure in Finland* under the Creative Commons Attribution License.

5.2 GENETIC ANCESTRY WITHIN FINLAND (STUDY II)

The fine-scale genetic structure, identified in Study I, focused on the individuals whose parents were born close to each other (under 80 km) and who thus were expected to have a relatively homogeneous background. However, this criterion of parents born close to each other excludes a majority of the FINRISK samples and describes the genetic structure only as present at the beginning of the 20th century. To get a more comprehensive understanding of the genetic background of the Finnish population, we developed and tested a framework to detect individual-level genetic ancestry within Finland in Study II.

5.2.1 FINE-SCALE REFERENCE GROUPS DETECT GENETIC ANCESTRY

Estimation of genetic ancestry builds on two resources: statistical methods and reference data. In Study II, we aimed to estimate the genetic ancestry within Finland by utilizing an existing statistical method implemented in the SOURCEFIND[52] program, and by characterizing easily interpretable and reliable reference groups. By following the procedure described in section 4.7.1, we identified three sets of reference groups that can be used for estimating ancestry with different levels of detail. The reference groups in refset 2 capture the genetic ancestry between the main division of East and West (Figure 16 A) and match well with the eastern and western populations identified in Study I. However, I note that the reference individuals located in Lapland are part of R2-East reference group in contrast to the previous results.

Furthermore, the reference groups in refsets 6 and 10 capture genetic ancestry with an exceptional level of detail: Refset 6 identifies two groups in the West and four groups in the East, while refset 10 identifies four additional groups, one in East Lapland and three along the genetic borderline between East and West. The pairwise-F_{ST} values (see manuscript of Study II) show that the smallest difference, 0.002, is between the groups called R10-Evacuated and R10-Central_Finland, while the largest difference, F_{ST} = 0.007, is between R10-Kainuu and R10-Bothnia, corresponding well to the F_{ST}-values that were seen between the fine-scale populations in Study I. While these 10 groups do not exactly match all the 17 fine-scale populations identified in Study I, they represent the geographic regions in Finland well and show sufficient genetic differentiation to serve as reference groups.

To test whether the identified reference groups can reliably detect ancestry, we simulated offspring using real-world data and compared the ancestry estimates of our framework to the geographic origin of the true ancestors of the simulated individuals. The simulations were conducted in two scenarios where the ancestors of the simulated offspring originated: 1) broadly from the East and West; and 2) from seven more detailed regions (see Figure 12) that approximately match the locations of the reference groups in refset 6 and some in refset 10.

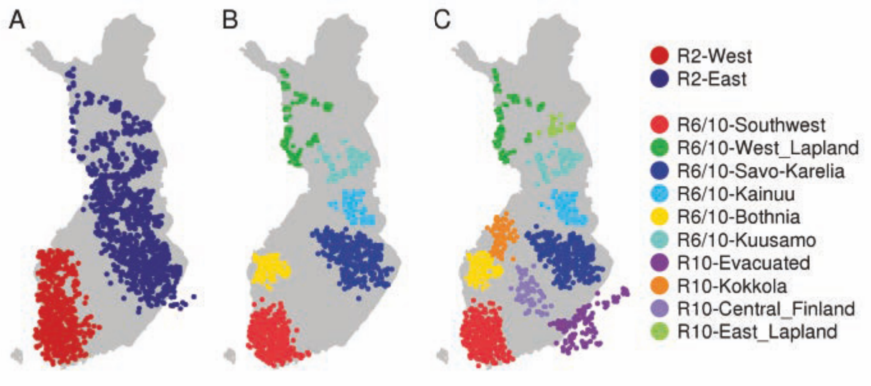


Figure 16. The geographic location of the reference groups used to detect genetic ancestry within Finland in A) refset 2, B) refset 6, and C) refset 10. The legend on the right shows the names of the reference groups. The figure is reprinted from the manuscript Kerminen *et al.* (2020) *Changes in the fine-scale genetic structure of Finland through the 20th century*.

The first simulation scenario, with ancestors from A-East and A-West, demonstrated that the refset 2 can accurately detect one ancestor coming from either of the populations 4 generations back. This result suggest that we can reliably estimate ancestry proportions that are 6% or larger with refset 2. The second simulation scenario, with the refset 6 and refset 10, demonstrated that we can accurately estimate the main source of ancestry from any region 3 generations back but the minor source of ancestry is underestimated if it comes from 2 or more generations back and the correct source of the minor ancestry is difficult to identify from 3 or more generations back. We should keep in mind that, as these simulations were based on real-world data, we do not know the correct genetic ancestry profile and the expected ancestry is based on the geographic origin of the ancestors.

5.2.2 CHANGES IN GENETIC ANCESTRY IN THE 20TH CENTURY

To characterize the admixture spectrum of genetic ancestry in Finland, we applied the framework of genetic ancestry estimation to 18,463 individuals from the FINRISK Study. These individuals were born all around Finland between 1923 and 1987 and allowed us to examine the changes in the genetic structure through a major part of the 20th century.

We grouped the 18,463 individuals into 12 groups based on their birth region, shown in Figure 17 and Figure 18, and averaged the individuals' ancestry estimates stratified by their birth years using a local regression (LOESS). Figure 17 shows the changes in the proportion of ancestry from R-West and R-East per region. Unsurprisingly, the western ancestry dominates the southwestern regions, SOF, SWF, TAV and OST, and the eastern ancestry dominates the regions of LAP, KAI, NOS, NSA, NKA, SSA, and SKA during the whole period. However, we detect significant changes in the proportions over time. The most dramatic change happened in CNF, where the eastern ancestry has increased and displaced the western ancestry as the largest source in the 1970s. Smaller changes were also detected, for example, in SOF and TAV, where the western component has decreased over 20 percentage points between 1930 and 1980. In OST, NOS, KAI, LAP, and SKA, the ancestry proportion has remained fairly constant with only some small fluctuation.

In Figure 18, we illustrate the changes in the genetic ancestry using the reference groups in refset 10. The results using refset 6 closely reflect the results of refset 10 and are shown in Study II. Overall, we detect similar behavior as with refset 2: the dominant ancestry originates from the nearest reference group and its proportion is often decreasing with time, suggesting that the genetic ancestry is diversifying with time. Interestingly, we also detect a few rapid changes that, in some cases, can be dated almost within the precision of one year. For example, in TAV and SWF, we detect a sharp increase in the R10-Evacuated ancestry (purple) in 1940. The increase in this ancestry source can be detected in all other regions, except in SKA and KAI. Table 3 compares the proportions of WWII evacuees in Finland, based on Paukkunen (1989)[127], to the genetic ancestry estimates in 1950 and shows that the genetic estimates are slightly higher than the recorded proportions of evacuees.

Table 3. The proportion of Karelian evacuees (from CKA) and the estimated genetic ancestry proportion of R10-Evacuated reference group in different regions in the year 1950. Comparisons with other regions was not feasible due to regional changes since 1950. The table is adapted from the manuscript Kerminen *et al.* (2020) *Annual changes in the fine-scale genetic structure of Finland through the 20th century*.

Region	Proportion of evacuees	Genetic ancestry estimate (95% CI)	
SOF	0.11	0.19	(0.18-0.21)
SWF	0.09	0.11	(0.10-0.12)
TAV	0.12	0.17	(0.15-0.20)
OST	0.05	0.07	(0.04-0.10)
NOS+KAI	0.03	0.07	(0.07-0.08)
LAP	0.04	0.08	(0.07-0.09)

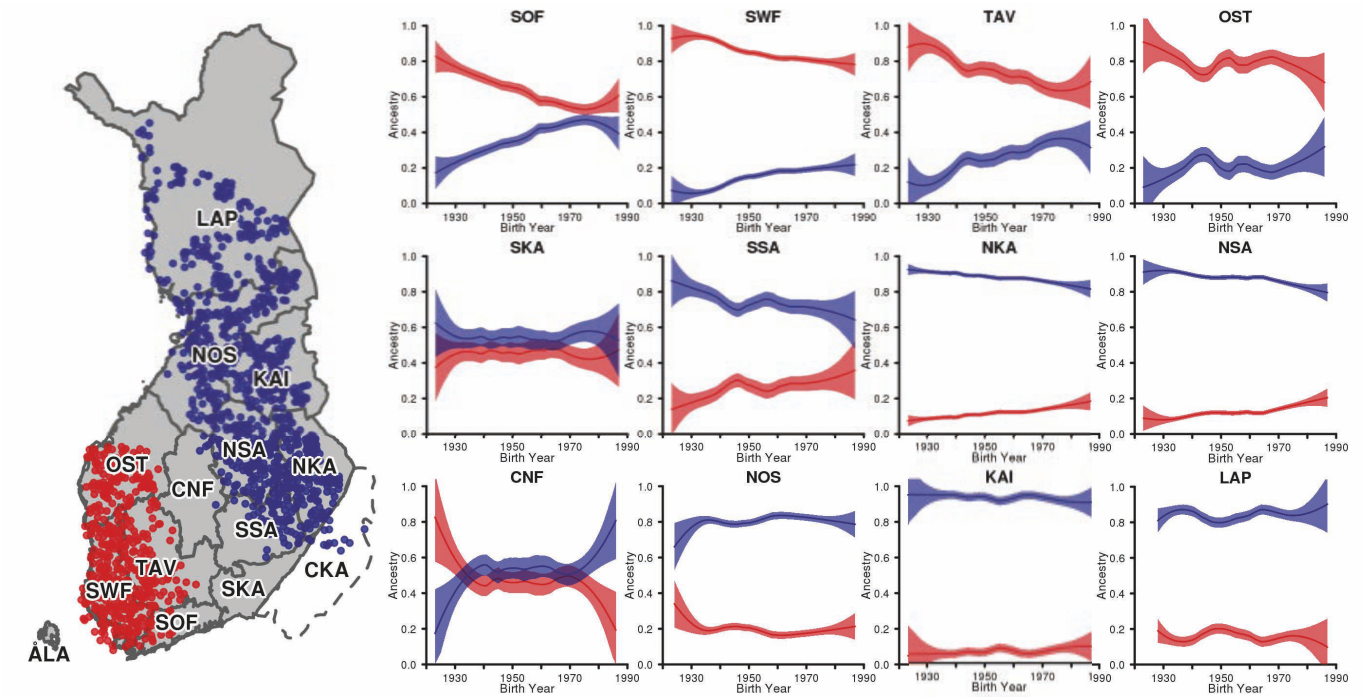


Figure 17. Changes in the genetic structure using level-2 reference groups from East and West. Map on the left shows the reference groups (red and blue dots) and the birth regions. The curves demonstrate the development in the genetic ancestry for the newborns in each region with 95% confidence intervals around the curves. The figure is reprinted from the manuscript Kerminen *et al.* (2020) *Changes in the fine-scale genetic structure of Finland through the 20th century*.

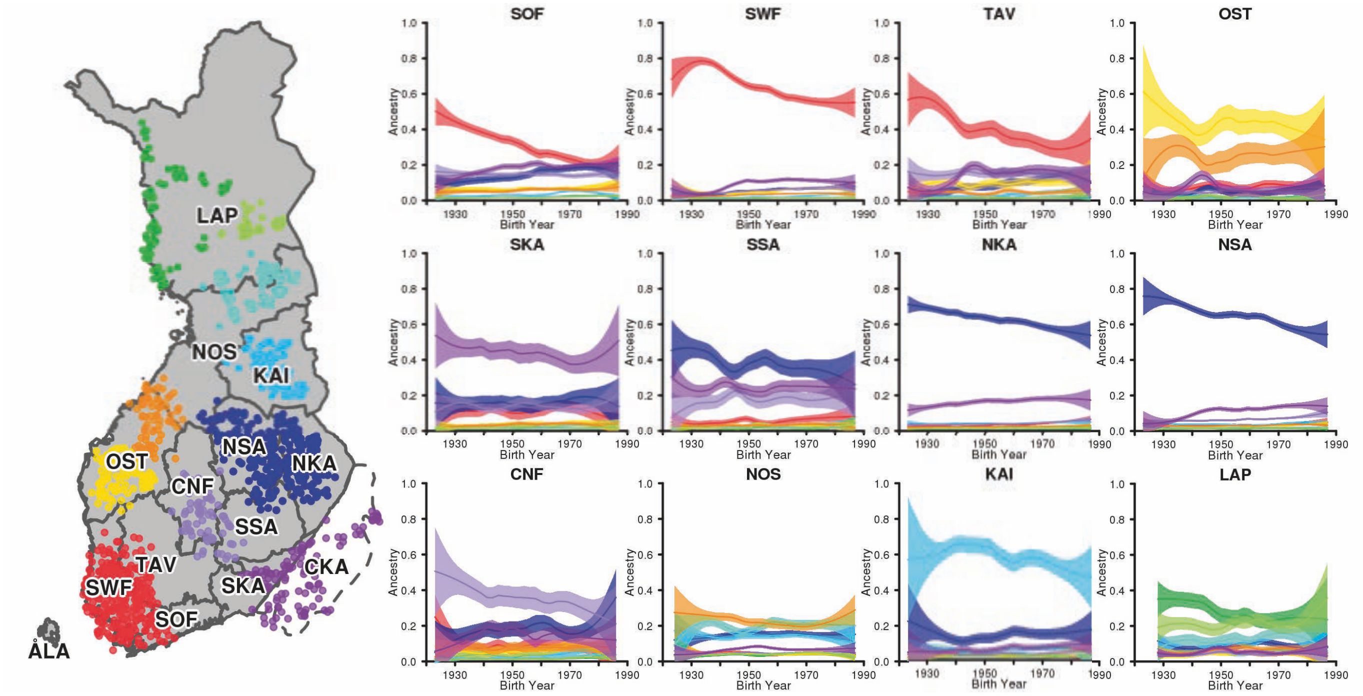


Figure 18. Changes in the genetic structure using level-10 reference groups. Map on the left shows the location of the reference groups and the birth regions. The curves demonstrate the development in the genetic ancestry of the newborns in each region with 95% confidence intervals around the curves. The figure is reprinted from the manuscript Kerminen *et al.* (2020) *Changes in the fine-scale genetic structure of Finland through the 20th century.*

5.3 GENETIC STRUCTURE AND POLYGENIC SCORES (STUDY III)

A polygenic score (PS) is a quantitative measure that summarizes the set of individual's genotypes after weighting them with the corresponding effect estimate from a GWAS. If we knew exactly which variants contribute to the disease or trait of interest, we could use polygenic scores to predict the differences also between individuals from different populations. However, the current polygenic scores can integrate information from tens of thousands to millions of variants, and thus even a tiny bias in the effect estimates of the underlying data can lead to flawed interpretations. It is especially critical to understand whether such biases exist, whether they amplify the differences in polygenic prediction between populations, and further affect the individual risk estimation. In Study III, we assessed the geographic variation and biases related to genetic structure in polygenic scores between the eastern and western populations in Finland.

5.3.1 POLYGENIC SCORES SHOW GEOGRAPHIC VARIATION

To evaluate the geographic variation in PS in Finland, we built polygenic scores for 5 diseases: coronary artery disease (CAD), rheumatoid arthritis (RA), Crohn's diseases (CD), ulcerative colitis (UC), and schizophrenia (SCZ), along with 3 quantitative traits: height (HG), body mass index (BMI) and waist-hip ratio (WHR), based on large international GWAS meta-analyses. We calculated these PS for 2,376 individuals from FINRISK. Then, by utilizing the geographic location information of the individuals, we generated risk maps that described the geographic distribution of the polygenic scores (Figure 19). These maps showed geographic differences for the PS of CAD, RA, SCZ, BMI, WHR and HG in East-West direction (statistics shown in Study III). In South-North direction, there were significant differences only for HG and WHR. PS of CD and UC did not show any significant geographic differences.

As most of the maps showed suspiciously similar patterns to the main genetic structure in Finland, we took a closer look at the observed differences in the quantitative traits. Especially for HG, the difference between East and West was peculiarly large, over 1.5 SD-units. In our sample, Eastern Finns were observed to be 1.7 cm shorter on average than Western Finns. When the 1.5 SD-unit difference in PS of HG was translated into a natural scale, it corresponded to 3.5 cm difference dramatically overestimating the observed phenotypic difference, especially when considering that the PS of HG explained only 14% of the variance in height. Similarly, the differences in PS of BMI and WHR showed exaggerated differences when compared to their phenotypic counterparts.

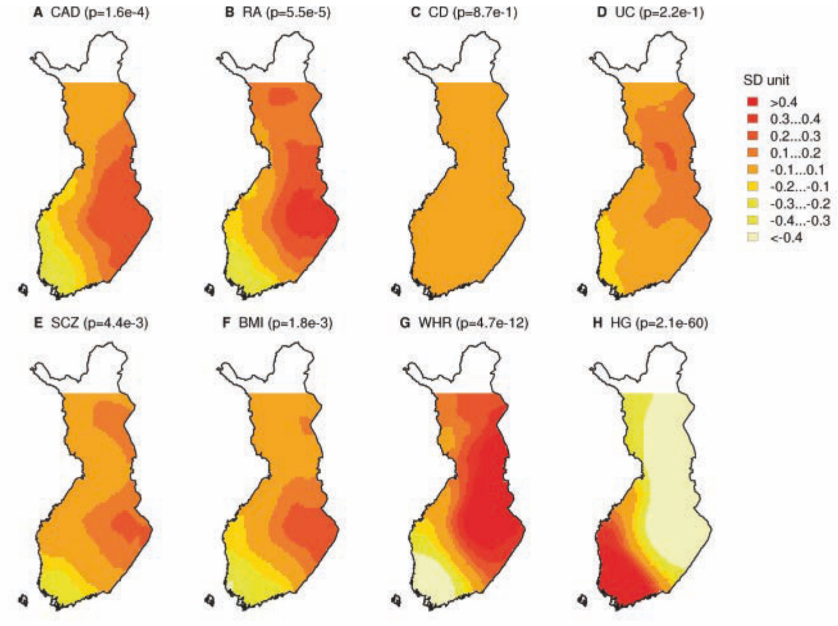


Figure 19. Geographic distribution of polygenic scores (PS) of complex diseases and traits in Finland. The scale corresponds to the standard deviation units of each PS distribution. The values above the maps correspond to the p-value of PS association with longitude. A) CAD = Coronary artery disease, B) RA = Rheumatoid arthritis, C) CD = Crohn's disease, D) UC = Ulcerative colitis, E) SCZ = Schizophrenia, F) BMI = Body mass index, G) WHR = Waist-hip ratio, H) HG = Height. The figure is reproduced from AJHG 104(6), Kerminen *et al.* (2019) *Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Trait in Finland* with permission from the publisher.

5.3.2 GENETIC STRUCTURE CAN BIAS POLYGENIC PREDICTION

Polygenic scores have been shown to exaggerate differences between genetically distant, continental populations such as African, Asian and European populations[3, 120]. The above results caused concerns about whether similar biases can arise even within more homogeneous populations, such as between Eastern and Western Finns. We assessed possible biases in our PS using adult height as a model trait. We compared three polygenic scores, built based on different summary statistics (GIANT, UKBB, FINRISK), and observed considerable differences in their predicted height difference between Eastern and Western Finland. While UKBB predicted only a 0.6 cm difference, FINRISK predicted a 1.7 cm difference and GIANT over a 3.5 cm difference. Furthermore, we compared additional polygenic scores utilizing different p-value thresholds and number of variants, and we were able to show that, especially GIANT-PS accumulated differences between East and West when the number of variants in PS increased. These results suggested that the existing summary statistics from large GWAS can contain subtle biases that can lead to unrealistically large PS differences between closely related populations.

To easily identify such biases in the summary statistics, we developed a new approach. The approach built additional PS using a random set of weakly associated variants from the original summary statistics and compared the accumulation of difference between populations when the number of variants was increased. We applied the approach to all eight PS and Figure 20 shows bias accumulation in all quantitative traits, HG, BMI, and WHR, that were based on the summary statistics from the GIANT consortium (2015). In addition, a bias accumulation was detected for summary statistics of CAD but not for other diseases in this analysis.

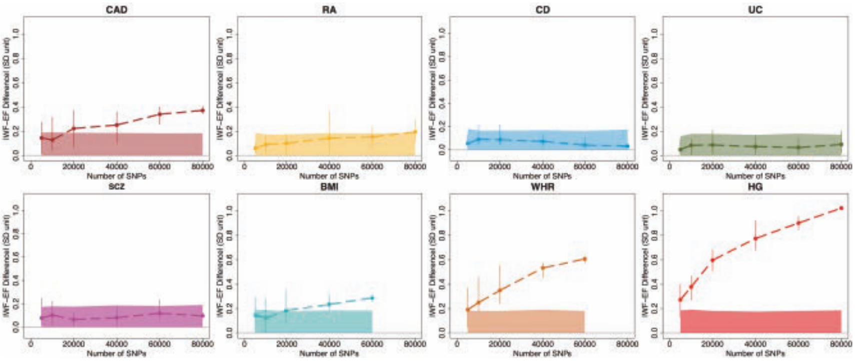


Figure 20. Accumulation of differences between Eastern Finland (EF) and Western Finland (WF) in polygenic scores (PS) built on weakly associated variants. The dots represent the absolute value of the PS difference between EF and WF with different numbers of independent, weakly associated variants (mean and the range of 10 scores are shown). The solid region shows the 95% probability interval above which PS are interpreted as showing a bias accumulation. The figure is reproduced from AJHG 104(6), Kerminen *et al.* (2019) *Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Trait in Finland* with permission from the publisher.

6 DISCUSSION

The genetic structure of a population is routinely controlled as a confounder in genome-wide association studies and the growing interest to utilize polygenic scores, derived from those analyses, has forced the genetic community to better understand the role of genetic structure in current analyses. In Finland, a large-scale biobank study, the FinnGen Project, is currently analyzing hundreds of thousands of samples and aims to translate the genomic information into health-care solutions. To ensure the success and equally distributed benefits between genetically diverse populations, a thorough understanding of genetic structure in Finland is required. This thesis presents the fine-scale genetic structure in Finland, illustrates the regional changes in the genetic structure during the 20th century, and underlines a comprehensive understanding of the connection between the genetic structure and polygenic scores.

6.1 GENETIC STRUCTURE IN FINLAND

In the real-world, natural populations are rarely discrete but show continuous genetic variation across geographic regions. Here, in Study I, the fine-scale genetic structure of Finland was modeled using only individuals whose parents were born close to each other and by utilizing a clustering-based FineSTRUCTURE program. Such an approach limits our understanding of the real-world genetic structure, first, by excluding a large part of the sample whose parents are not from the same geographic background, and second, by giving an unrealistically discrete picture of the genetic populations highlighting boundaries between the populations. However, the clustering-based approach enables clear visualization of results, and, together with admixture-type analyses, can simplify the interpretation of results. In the following, the fine-scale genetic structure of Finland is discussed as it appeared in the results, while acknowledging that the underlying structure is more continuous and diverse than the results are able to easily summarize.

Eastern and Western populations
The genetic division into Eastern and Western Finland has been acknowledged for decades[110, 139, 140, 147-149, 153, 154, 161] but a purely

genetically defined location of that borderline between East and West has been missing. Here, by utilizing genome-wide SNP data, a geographically comprehensive sample and modern haplotype-based methods, we were able to accurately locate the borderline between Eastern and Western populations. Comparing to the prehistoric, historic, and dialect regions demonstrated several similarities to the genetic borderline, and in Study I, the border of the Treaty of Nöteborg was highlighted as a particularly close match to the genetic borderline. However, the genetic borderline shares features with most East-West borderlines, from the Corded Ware culture in the Stone Age to the main dialect regions today. Therefore, it is unlikely that the genetic borderline we see today is a result of a single historical event but rather a result of a long-term process with consecutive influences from East and West that have been reinforced by the population growth. Considering also the results presented in Study II, it should also be emphasized that the genetic borderline between East and West is not a strict division into two separate populations, but a gradual transition between the two populations.

The analyses, repeated with different sample sizes, demonstrated that the individuals in Lapland were clustered into West with the smaller sample sizes and into East with the larger sample. This could be explained by the fact that, in the smaller data sets, Lapland had a limited and western-biased sample. However, the northern Finns also showed rather independent genetic characteristics as they were the next to separate after the East-West split in both PCA and FineSTRUCTURE analyses, and they have been shown to exhibit strong internal structure[160]. Thus, it can be artificial to model the Northern Finns as only Eastern or only Western Finns, but instead, their own particular genetic characteristics should be considered, also including the possible ancestry from the Sami. Thus, the clustering of Lapland is heavily affected by the relative proportions of the ancestry of the sampled individuals, and the region would benefit from a larger haplotype-based study of its own.

Fine-scale genetic populations

The identification of fine-scale populations revealed dozens of previously unseen genetic populations from which 17 higher-level populations were geographically tightly clustered and covered mainland Finland evenly, and agree well with the results of other haplotype-based analyses[81, 163].

The result reflects the known population history of Finland as a relatively young population with a long-standing isolation-by-distance within the country. This observation is emphasized when our results are compared to the similar analyses in other countries, e.g., in the UK[53] and in Denmark[192], where notable genetic homogeneity was found. For example, while the first FineSTRUCTURE analysis from the UK[53] identified 53 fine-scale genetic populations, it still clustered almost half of the individuals into one population covering most parts of England, excluding only the southwestern parts. The difference to our results is striking when it is considered that, in the UK study, they examined the genetic structure of individuals whose all four grandparents were born near each other, while in our analysis, the genetic structure was examined only at the level of the parents.

In the analyses of Middle and Southern European countries, such as in France[60], Spain[61] and the Netherlands[193], the fine-scale population structure has been partly explained by physical barriers such as rivers and hills. In our analyses, we did not recognize similar geographic explanations for the fine-scale structure, which highlights the importance of the isolation-by-distance phenomenon. Contrary to the European countries mentioned above, in Finland the rivers and lakes have most likely been important passageways connecting regions dominated by forests and wilderness. For example, the population P12 in Figure 15 seems to be distributed along the river Ii. However, the resolution of this study is not sufficient for more than a speculative interpretation of the role of geographic barriers in Finland because the location information was at the municipality level only.

While the fine-scale genetic structure presented here provides a good foundation for the ongoing genetic studies, the increasing sample size in the future will provide new opportunities to detect even more details. However, I consider it more interesting to study the current results together with genetic data from neighboring populations (Sweden, Estonia, Russia), as is already studied in the populations of Estonia[194], France[60] and the Neatherland[62], to gain a broader context on the relationships of the populations within and outside of Finland.

6.2 CHANGES IN THE REGIONAL GENETIC STRUCTURE

Genetic structure is often seen as a static snapshot of the sample at hand. Nonetheless, gene flow due to migration can shape the genetic composition of a population substantially and rapidly. An evident example is the migration to the Americas starting from the 1500s, both due to the forced slave trade and voluntary immigration across the globe that has transformed the native American gene pool into a heterogeneous mixture of global ancestries over the last five hundred years[48, 49, 195]. In Finland, the migration and admixture patterns have been much subtler but, by combining modern haplotype-based methods with birth-year and birth-place information, we were able to track detailed changes in the regional genetic structure in Finland within the last century. While our results mirrored closely the recent population history, they also demonstrated that the genetic composition on small geographic areas can change surprisingly rapidly.

The changes in the regional genetic structure within Finland demonstrated a general trend of a decrease in the dominant ancestry component. With refset 2, this trend was strongest in the central and southern parts of Finland and weakest in the northern and eastern regions, suggesting that the direction of the gene flow during the 20th century has headed from East to West. While a confirmation for this observation would require a detailed investigation of the Finnish migration records, it is at least supported by the fact that the four largest cities, Helsinki (including the entire capital region), Tampere, Oulu, and Turku, are located in the western parts of the country. Along with the decrease in the dominant ancestry component, the proportions of other ancestry components increased steadily, implying that the genetic ancestry profiles have diversified in many regions during the century. This diversification is assumed to have continued along with the expansion of urbanization and immigration towards the end of the 20th century[196].

On top of the general trends, the ancestry curves revealed some very rapid changes that, in some cases, could be detected with an annual accuracy. The most prominent example was the increase in the Evacuated ancestry in the 1940s, which matched well with the actual migration of the WWII evacuees from the ceded Karelian regions. The increase in the Evacuated component was the strongest in the southern, western, and central parts of Finland, and weakest in northern Finland, which was consistent with the historical records[127]. Although there seemed to be a slight overestimation in the genetic ancestry of R10-Evacuated reference group. In the regions genetically most closely related to the evacuated Karelian people, such as in South and North Karelia, the estimates of war-related migration could not be reliably inferred with the genetic approach presented here. Together, the results concretely demonstrate the power of genetics to reveal historical events and argue that a collection of older and/or more diverse samples may uncover already forgotten events. As the genetic structure continues to develop, a younger sample with individuals born in the 1990s and after would provide an exciting extension for the period covered here and offer a glimpse of today's the genetic structure.

6.3 GENETIC ANCESTRY ESTIMATION

Here, the genetic ancestry was estimated as a genetic similarity to predefined reference groups. The main limitation of such an approach is that the results are always heavily dependent on the references used and can even ignore major sources of ancestries that are not included in the reference set. In Study II, the reference groups covered most geographic regions of Finland well, and in refset 10, only regions near the cities of Helsinki, in Southern Finland, and Oulu, in North Ostrobothnia, were without a representative reference population. Indeed, the regions of Southern Finland and North Ostrobothnia, lacking a clear reference population of their own, showed considerably diverse ancestry profiles. In addition, Study II did not include reference samples from minority groups within Finland, e.g., the Sami people, and hence additional studies would be needed to examine the contribution of minority groups to the Finnish gene pool. Also, the future studies combining modern and ancient data both from Finland and abroad would shed light on the population history more broadly.

Despite the possible limitations, the simulations with real-world data showed that the given references groups were able to accurately estimate the expected geographic origin. With refset 2, the ancestry from East or West was confidently detected, if 1 out of 16 grandparents (corresponding to around 6% of ancestry) 4 generation back originated from that reference group. With refset 6 and 10, the detectability was weaker but permitted identifying the source of ancestry coming from at least 1 out of the 4 grandparents.

Estimation of an individual's genetic ancestry is of great interest among the general public, but it is also useful in genetic studies of disease and traits. A recent study has, for example, shown that ancestry estimation can improve the genetic prediction in recently admixed individuals by modeling partial polygenic scores based on the individual's local ancestry[121]. Thus, in addition to detecting the historical events discussed above, the ancestry estimation presented here could be employed to improve the genetic risk estimation in Finland.

6.4 POLYGENIC SCORES IN FINLAND

Polygenic scores (PSs) are currently extensively used in genetic studies and their utilization in health care settings is anticipated[113]. PSs could be used to better identify individuals at high risk but also to motivate patients toward better lifestyle choices. Nonetheless, PSs are criticized for being weak in individual risk assessment. Also, PSs have been shown to be a poor tool for quantifying genetic risk differences between distant populations, such as Europeans and Africans[3, 120]. In Study III, we assessed the polygenic score variation within Finland, and the results demonstrated that similar problems of comparing genetic risk differences between populations can manifest even between much more closely related populations.

By focusing on height as a model trait, we showed that PSs derived from different summary statistics can predict varying differences between the Finnish subpopulations. In particular, the PS derived from the results of the GIANT consortium[115] showed the most exaggerated differences. A more detailed examination revealed a substantial accumulation of differences along with the increasing number of variants in GIANT-PS, suggesting a small but consistent directional bias in effect estimates that aligns with the main genetic structure in Finland. Further analyses showed that an exclusion of the Finnish samples from the GIANT meta-analysis approximately halved the differences but a considerable bias still remained. Two other studies have also reported serious biases in the same GIANT data[116, 117], but the exact mechanism of such a bias still remains unexplained. Our approach to detect a bias accumulation in PS between two populations revealed that also waist-hip ratio and body mass index from the GIANT consortium and coronary artery disease from CARDIoGRAMplusC4D[173] suffered from considerable biases. Nonetheless, it should be noted that possible new analyses of these consortiums and traits may not entail such biases and the existing and future summary statistic should be separately tested for possible biases.

Although PS of coronary artery disease and the quantitative traits showed considerable bias accumulation, the same was not observed for the other diseases even though some of them showed geographic differences in their disease incidence. PS of schizophrenia did not show a considerable bias accumulation but it showed some geographic variation. The extensive prevalence information of schizophrenia[197-202] in Finland shows the lowest rates in the southwestern corner of Finland consistent with the PS distribution, suggesting that population-genetic factors behind the regional variation in schizophrenia prevalence should not be excluded. The regional prevalence and incidence information for the other diseases is however more limited. For rheumatoid arthritis, PS again showed higher risk in Eastern than in Western Finland, and a study[203] looking at the incidence of the disease itself reports higher incidence rates in North Karelia (in the East) and lower in Ostrobothnia (in the West), but does not include the northern or southwestern parts of Finland in the analysis. For ulcerative colitis and Crohn’s disease, the geographic differences in PS were not significant and only a subtle difference between the South and the North[181], and between urban and rural regions[204] have been reported before. Together these examples indicate both the challenge but also the potential of polygenic scores in helping us understand the role of genetic structure in the regional differences of diseases and traits. Further studies combining comprehensive prevalence information with genetic and environmental data, as well as unambiguous methods for translating PS difference into a prevalence scale, are needed to reveal the role of genetic factors in regional heath differences.

6.5 POPULATION GENETICS AND THE GENERAL PUBLIC

Population genetics is of interest to the general public, because it can provide new and sometimes unexpected answers to the fundamental questions of our origin and identity. Indeed, millions of people have already taken commercial ancestry tests provided by private companies[205]. People are often using these tests to discover their genetic origin and for tracking unknown relatives to build personal genealogies. However, the companies are also providing information about customers' genetic health and disease risk, and both the private and public large-scale data repositories are being collected and further used both for commercial and scientific purposes[206]. Therefore, the interface of population genetics and the general public is tightly linked to the current challenges of human genetics in general. Three main challenges are summarized below.

First, privacy and data protection have become a central part in genetics research. Similar to any personal data, genetic data are sensitive information, potentially identifying individuals. However, a special feature of genetic data is that they may also lead to identifying relatives. Along with the growing number and size of public and private databases, also the risks of misuse and information leakage increase. To maintain the trust and the personal safety of both patients and customers and their relatives, the field of human genetics has worked—and needs to continue working hard—to develop trustworthy and transparent procedures for data management.

Second, ancestry and other genetic testing can reveal unexpected results that are not always considered positive and thus can cause serious distress. For example, the tests can unintentionally reveal that the expected father is not the biological father or that an individual has a substantial predisposition to a disease without any effective treatment options. At the same time, such tests also reveal information about the relatives of the tested person even though the relatives have not themselves consented to any tests. Thus, a careful consideration of the possible benefits contrasted with the possible disadvantages should be done both by the professionals designing the tests and by the target audience.

Third, genetics has been and is being misused to justify discrimination, stigmatization, and even physical harm towards specific groups of people, often minorities. The misuse often aims to explain physical or behavior differences between groups but, as has been demonstrated by multiple studies[3, 119, 120, 207] and by this thesis, it is extremely challenging to robustly link phenotypic differences and genetic differences between populations. This is because other factors, such as geographic location and related environmental factors—such as income and education, as well as the possible technical problems—can confound the results. Additionally, population genetics is misused to pseudo-scientifically justify a hard classification of people. However, such hard classification is rarely scientifically motivated because, while showing geographic patterns, genetic variation is not a discrete variable but a continuum[208, 209]. While this continuum could be artificially split, as is also done in academic studies for practical reasons, there is no naturally motivated break or split points between populations. This is one reason why the results of this thesis do not provide a definition of, e.g., a Karelian individual, an Eastern Finn, or a Finnish individual. To ensure the understanding of these issues among the general public, the research community needs to educate and openly communicate about their research.

As part of the studies presented in this thesis, we have communicated our science to the general public to promote open research and learning, both of which are also part of the strategic plan of the University of Helsinki 2021-2030[210]. In addition to the articles published as open access, we published (as a part of the studies I and II) web pages where the results of our research could be examined. The web page for Study I (https://www.fimm.fi/en/research/projects/finnpopgen) received 20,000 visitors within a month after the study was published and attracted also media attention from large Finnish newspapers such as Helsingin Sanomat[211] and Ilta-Sanomat[212]. I believe that this demonstrates the great interest of the general public towards population genetics, which will hopefully translate into active participation in future research.

7 CONCLUSIONS

Genetic studies have entered the biobank era, in which hundreds of thousands of samples are automatically analyzed and the utilization of the results in a health-care setting is anticipated. Polygenic scores are a state-of-the-art tool for translating the genetic information into interpretable measures. However, polygenic scores suffer from issues closely related to population genetics. The lack of diversity in GWAS studies, poor translation between populations as well as biased effect estimates can endanger the quality of these tools and hinder their adoption in health care. A thorough understanding of population genetics and genetic structure is a key component for solving these challenges. At the same time, population genetic data can provide us a fascinating insight into the population history.

This doctoral thesis characterized novel details about the genetic structure of Finland during the 20th century and argues that the geographic variation and bias in polygenic scores of complex diseases are often tightly connected to this structure in Finland.

The results of the genetic structure provide fine-scale details in terms of both the geographic and temporal structure. The work identified 17 geographically clustered and robust populations within Finland and mapped the changes in the regional structure with annual accuracy. The results brought our understanding of Finland's genetic structure into the age of haplotype-based methods. The work also serves as an example of estimating genetic ancestry within a single country as well as a platform for biobank-scale ancestry estimation in Finland. In the future, the results can be used as a basis for improved control of genetic structure in genome-wide association studies. However, to gain richer insight into the population history further back in time, comparative studies including a wide range of neighboring populations and ancient DNA samples are needed.

Moreover, the work demonstrated that the geographic variation in polygenic scores, at least partly, arises from population-genetic biases that are not yet fully understood. Thus, polygenic scores should not be used to explain health differences between populations before the exact mechanism of population-genetic biases in polygenic scores are understood. Nevertheless, the results do not exclude genetic factors as a component underlying the geographic variation in diseases and traits. To understand and overcome the challenges of polygenic risk prediction for the benefit of human health, the increase in sample sizes and methodological improvements, alone, are not sufficient. For efforts to truly overcome these challenges, a thorough understanding on the general genetic variation in human populations is needed.

ACKNOWLEDGEMENTS

This doctoral thesis was carried out at the Institute for Molecular Medicine Finland (FIMM) of the University of Helsinki. I want to acknowledge the directors of FIMM, Mark J. Daly, Jaakko Kaprio, and Olli Kallioniemi for providing a high-quality research facility as well as an inspiring and supportive work environment during my doctoral studies.

The work on this thesis has been financially supported by the Doctoral Program in Population Health of the University of Helsinki, the Academy of Finland, The Sigrid Juselius Foundation, the Emil Aaltonen Foundation, and the Finnish Concordia Fund. I want to thank these institutions and foundations for their generous support. I also want to thank the student organization of Varsinaissuomalainen osakunta of the University of Helsinki for their stipend.

For my supervisors, Matti Pirinen and Samuli Ripatti, I want to express my deepest gratitude for guiding me through what has been my doctoral dream. Matti, thank you for your endless patience, attention to detail, and devotion to science. You taught me the value of careful and disciplined work, but also reminded me that it is important to work due to ones' own interest. Samuli, thank you for your bright and encouraging attitude as well as your wide-ranging insights. Your guidance helped me step back from the analyses and to move forward with my project. (In addition, thank you Samuli for accepting the role of Custos at my defense!) Together, you two have been the best possible combination of supervisors I could have hoped for.

For the opponent and reviewers, I want to express my sincere appreciation for taking the time to review my thesis. Thank you Tuuli Lappalainen, for accepting the request to serve as the opponent in these exceptional circumstances. Thank you Outi Savolainen and Kristiina Tambets, for pre-examining the thesis and for your valuable comments on it.

I want to acknowledge the members of my thesis committee, Antti Sajantila and Päivi Onkamo. Over the years, you have challenged me to step back and look at my work from new perspectives. Also, a special

thanks go to Päivi for being a part of my university studies all the way from the bachelor's to the master's and finally to the doctoral degree.

As we all know, nothing in life or in science is done alone, and this thesis is not an exception. I have had the great privilege to work in this project together with talented and trustworthy co-authors. Thank you, Aki Havulinna, Alicia Martin, Garrett Hellenthal, Sanni Ruotsalainen, Jukka Koskela, Ida Surakka, Antti-Pekka Sarin, Pekka Jousilahti, Markus Perola, Aarno Palotie, Mark Daly and Veikko Salomaa. Without your contributions on data, samples, methods, analyses, and your profound expertise in the field, this work would not have been possible. Thank you, Nicola Cerioli, Darius Pacauskas and Rupesh Vyas for implementing a beautiful website for the general public to examine the results of Study II.

I am also greatly indebted to the whole scientific community who has supported me along the way. First, thank you Iiris Hovatta for guiding me toward FIMM and to the Ripatti and Pirinen Groups to work on this project in the first place. Thank you also, Lisa Muszynski, for revising the language of my thesis. Second, a big thank you goes to the whole FIMM community. Thank you FIMM coordinators and administrative staff, Eiri, Emilia, Huei-yi, Sanni, Sari, Susanna, and Ulla, for keeping things rolling. A special thank you to Mari Kaunisto, for offering unique and surprising opportunities to communicate our science to a wider audience. Thank you FIMM IT personnel, especially Olle and Timo, for your practical help in keeping my analyses running. Thank you FIMM PhD Student and PostDoc Council and the Fellow members, Lea, Dimitrios, Pu, Andrew, Juho, Kalle, Emma, Jennifer, Lassi and Alok, for balancing the daily life of our community with cheerful activities. Finally, my warmest thanks go to my dear colleagues. Thank you, Elina, Elmo, Hande, Heidi, Joel, Juulia, Lea, Linda, Mari, Meri, Nella, Nina, Pietari, Pyry, Sanni and Shabbeer, for sharing your thoughts and workdays with me. Thank you, Aarno, Aki, Alyce, Andrea, Anni, Annina, Antti, Ari, Arto, Awaisa, Clara, Eija, Elisabeth, Emmi, Hanna, Hannele, Henrike, Himanshu, Ida, Jaakko K, Jaakko L, Jadwiga, Jake, Jarmo, Johanna, Juha, Jukka, Kalle, Kimmo, Leevi, Leif, Liisa, Luca, Maarit, Maija, Marita, Marjo, Mark, Mart, Mary Pat, Mattia, Mervi, Mikko, Mitja, Olli, Om, Paavo, Paula, Pietro, Priit, Risto, Rubina, Sakari, Susanna, Taru, Tiia, Timo, Tuomo, Vincent, Vishal, William, Xavier, Yu, Zuzanna, and many others whom I had the privilege to meet over the course of my project—the value of our daily coffee breaks, mutual helpfulness, and informal chats have been, as I have learned during the remote working period, irreplaceable for me.

I want to express my gratitude also to those who have supported me outside of the daily work. Thank you Riikka, Salla, Johanna, Anna, Katja, Eemeli, Fredu, and Mikko for friendship, amazing student life and PhD peer support.

To my family, Anne, Jyrki and Paula, thank you for your endless, unconditional love and support. You have given me strength to chase my dreams already well before they were on the horizon.

Johann, you know that the words mean more when I say them in Finnish: *Sinun apusi tällä matkalla on ollut laajempaa kuin kenenkään muun: Kiitos koodausvinkeistä, englannin kielen artikkeleista, harjoitusyleisönä olemisesta, tukevasta olkapäästä ja lämpimästä kodista. Kiitos, että olet ollut mukana matkan jokaisessa käänteessä.*

Sini Kerminen

December 2020

REFERENCES

1. Green RE, Krause J, Briggs AW, *et al.* A draft sequence of the Neandertal genome. *Science*. 2010; 328(5979):710-722.
2. Scutari M, Mackay I, Balding D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLoS Genet*. 2016; 12(9):e1006288.
3. Martin AR, Gignoux CR, Walters RK, *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet*. 2017; 100(4): 635-649.
4. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet*. 2012; 8(1):e1002453.
5. Freeman JL, Perry GH, Feuk L, *et al.* Copy number variation: new insights in genome diversity. *Genome Res*. 2006; 16(8):949-61.
6. Consortium TGP, Auton A, Brooks LD, *et al.* A global reference for human genetic variation. *Nature*. 1 2015; 526(7571):68-74.
7. Mallick S, Li H, Lipson M, *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. 2016; 538(7624):201-206.
8. Karczewski KJ, Weisburd B, Thomas B, *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017; 45(D1):D840-D845.
9. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020; 581(7809):434-443.
10. Glick BR, Pasternak JJ, Pattern CL. DNA Microarray Technology. In: Microbiology ASo, ed. *Molecular Biotechnology: Principles and applications of recombinant DNA*. 4 ed. ASM Press; 2010:155-160.
11. Scally A. The mutation rate in human evolution and demographic inference. *Curr Opin Genet Dev*. 2016; 41:36-43.
12. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet*. 1998; 63(3):861-9.
13. Kong A, Gudbjartsson DF, Sainz J, *et al.* A high-resolution recombination map of the human genome. *Nat Genet*. 2002; 31(3):241-7.
14. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*. 2006; 7(10):771-80.
15. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010; 26(22):2867-73.
16. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011; 88(1):76-82.
17. Reich DE, Cargill M, Bolk S, *et al.* Linkage disequilibrium in the human genome. *Nature*. 2001; 411(6834):199-204.

18. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet*. 2001; 29(2):229-32.
19. Gabriel SB, Schaffner SF, Nguyen H, *et al.* The structure of haplotype blocks in the human genome. *Science*. 2002; 296(5576):2225-9.
20. The International HapMap Consortium. The International HapMap Project. *Nature*. 2003; 426(6968):789-96.
21. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005; 437(7063):1299-320.
22. The International HapMap Consortium, Frazer KA, Ballinger DG, *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449(7164):851-61.
23. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*. 2015; 31(21):3555-7.
24. Hartl DL, Clark AG. *Principles of population genetics*. 4. ed. Sinauer Associates; 2007.
25. Coop G. *Population and quantitative genetics*. University of California, Davis; 2020.
26. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*. 2012; 336(6082):740-3.
27. Rosenberg NA, Pritchard JK, Weber JL, *et al.* Genetic structure of human populations. *Science*. 2002; 298(5602):2381-5.
28. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A*. 2005; 102(44):15942-7.
29. Jakobsson M, Scholz SW, Scheet P, *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*. 2008; 451(7181):998-1003.
30. Wang C, Zollner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet*. 2012; 8(8):e1002886.
31. Peter BM, Petkova D, Novembre J. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. *Mol Biol Evol*. 2020;37(4):943-951.
32. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science*. 1978; 201(4358):786-92.
33. Sokal RR. Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci U S A*. 1988; 85(5):1722-6.
34. Novembre J, Johnson T, Bryc K, *et al.* Genes mirror geography within Europe. *Nature*. 2008; 456(7218):98-101.
35. Lao O, Lu TT, Nothnagel M, *et al.* Correlation between genetic and geographic structure in Europe. *Curr Biol*. 2008; 18(16):1241-8.
36. McEvoy BP, Montgomery GW, McRae AF, *et al.* Geographical structure and differential natural selection among North European populations. *Genome Res*. 2009; 19(5):804-14.

37. Busby GBJ, Hellenthal G, Montinaro F, *et al.* The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape. *Curr Biol*. 2015; 25(21):2878.
38. Tishkoff SA, Reed FA, Friedlaender FR, *et al.* The genetic structure and history of Africans and African Americans. *Science*. 2009; 324(5930):1035-44.
39. Gurdasani D, Carstensen T, Tekola-Ayele F, *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature*. 2015; 517(7534):327-32.
40. van Dorp L, Balding D, Myers S, *et al.* Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. *PLoS Genet*. 2015; 11(8):e1005397.
41. Busby GB, Band G, Si Le Q, *et al.* Admixture into and within sub-Saharan Africa. *Elife*. 2016; 5.
42. Patin E, Lopez M, Grollemund R, *et al.* Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*. 2017; 356(6337):543-546.
43. Chaichoompu K, Abegaz F, Cavadas B, *et al.* A different view on fine-scale population structure in Western African populations. *Hum Genet*. 2020; 139(1):45-59.
44. Jeong C, Balanovsky O, Lukianova E, *et al.* The genetic history of admixture across inner Eurasia. *Nat Ecol Evol*. 2019; 3(6):966-976.
45. The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science*. 2009; 326(5959):1541-5.
46. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019; 576(7785):106-111.
47. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet*. 2015; 96(1):37-53.
48. Han E, Carbonetto P, Curtis RE, *et al.* Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat Commun*. 2017; 8:14238.
49. Dai CL, Vazifeh MM, Yeang CH, *et al.* Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. *Am J Hum Genet*. 2020; 106(3):371-388.
50. Ruiz-Linares A, Adhikari K, Acuna-Alonzo V, *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*. 2014; 10(9):e1004572.
51. Homburger JR, Moreno-Estrada A, Gignoux CR, *et al.* Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet*. 2015; 11(12):e1005602.
52. Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, *et al.* Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat Commun*. 19 2018; 9(1):5388.
53. Leslie S, Winney B, Hellenthal G, *et al.* The fine-scale genetic structure of the British population. *Nature*. 2015; 519(7543):309-314.

54. Gilbert E, O'Reilly S, Merrigan M, *et al.* The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Sci Rep*. 2017; 7(1):17199.
55. Gilbert E, O'Reilly S, Merrigan M, *et al.* The genetic landscape of Scotland and the Isles. *Proc Natl Acad Sci U S A*. 2019; 116(38):19064-19070.
56. Byrne RP, Martiniano R, Cassidy LM, *et al.* Insular Celtic population structure and genomic footprints of migration. *PLoS Genet*. 2018; 14(1):e1007152.
57. Takeuchi F, Katsuya T, Kimura R, *et al.* The fine-scale genetic structure and evolution of the Japanese population. *PLoS One*. 2017; 12(11):e0185487.
58. Yasuda J, Katsuoka F, Danjoh I, *et al.* Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku Medical Megabank Project. *BMC Genomics*. 2018; 19(1):551.
59. Sakaue S, Hirata J, Kanai M, *et al.* Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nat Commun*. 2020; 11(1):1569.
60. Saint Pierre A, Giemza J, Alves I, *et al.* The genetic history of France. *Eur J Hum Genet*. 2020; 28:853-865.
61. Bycroft C, Fernandez-Rozadilla C, Ruiz-Ponte C, *et al.* Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula. *Nat Commun*. 2019; 10(1):551.
62. Byrne RP, van Rheenen W, Project Min EALSGC, van den Berg LH, Veldink JH, McLaughlin RL. Dutch population structure across space, time and GWAS design. *Nat Commun*. 2020; 11(1):4556.
63. Bryc K, Auton A, Nelson MR, *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A*. 2010; 107(2):786-91.
64. Behar DM, Yunusbayev B, Metspalu M, *et al.* The genome-wide structure of the Jewish people. *Nature*. 2010; 466(7303):238-42.
65. Haber M, Gauguier D, Youhanna S, *et al.* Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet*. 2013; 9(2):e1003316.
66. Creanza N, Feldman MW. Worldwide genetic and cultural change in human evolution. *Curr Opin Genet Dev*. 2016; 41:85-92.
67. Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. The origins of lactase persistence in Europe. *PLoS Comput Biol*. 2009; 5(8):e1000491.
68. Haak W, Lazaridis I, Patterson N, *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. 2015; 522(7555):207-11.
69. Lazaridis I, Nadel D, Rollefson G, *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature*. 2016; 536(7617):419-24.
70. Wang CC, Reinhold S, Kalmykov A, *et al.* Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat Commun*. 2019; 10(1):590.

71. Saag L, Laneman M, Varul L, *et al.* The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr Biol*. 2019; 29(10):1701-1711 e16.
72. Lazaridis I, Patterson N, Mittnik A, *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. 2014; 513(7518):409-13.
73. Mathieson I, Scally A. What is ancestry? *PLoS Genet*. 2020; 16(3):e1008624.
74. Coop G. How many genetic ancestors do I have? *gcbias* blog. 2013. https://gcbias.org/2013/11/11/how-does-your-number-of-genetic-ancestors-grow-back-over-time/
75. Novembre J, Peter BM. Recent advances in the study of fine-scale population structure in humans. *Curr Opin Genet Dev*. 2016; 41:98-105.
76. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017; 95:1.22.1-1.22.23.
77. Wangkumhang P, Hellenthal G. Statistical methods for detecting admixture. *Curr Opin Genet Dev*. 2018; 53:121-127.
78. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; 155(2):945-59.
79. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 2003; 164(4):1567-87.
80. Gattepaille LM, Jakobsson M. Combining markers into haplotypes can improve population structure inference. *Genetics*. 2012; 190(1):159-74.
81. Martin AR, Karczewski KJ, Kerminen S, *et al.* Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. *Am J Hum Genet*. 2018; 102(5):760-775.
82. Wright S. The genetical structure of populations. *Ann Eugen*. 1951; 15(4):323-54.
83. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. *Genome Res*. 2013; 23(9):1514-21.
84. The 1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491(7422):56-65.
85. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006; 2(12):e190.
86. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8):904-9.
87. *PLINK 1.9*. www.cog-genomics.org/plink/1.9/
88. Galinsky KJ, Bhatia G, Loh PR, *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet*. 2016; 98(3):456-472.

89. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol*. 2015; 39(4):276-93.
90. Hellenthal G, Busby GBJ, Band G, *et al.* A genetic atlas of human admixture history. *Science*. 2014; 343(6172):747-751.
91. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. 2003; 165(4):2213-33.
92. The GWAS Catalog. The NHGRI-EBI Catalog of published genome-wide association studies. The European Bioinformatics Institute (EMBL-EBI) and National Human Genome Research Institute (NHGRI). Accessed 24.4.2020, https://www.ebi.ac.uk/gwas/
93. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012; 90(1):7-24.
94. Visscher PM, Wray NR, Zhang Q, *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. 2017; 101(1):5-22.
95. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet*. 2003; 361(9357):598-604.
96. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004; 36(5):512-7.
97. Campbell CD, Ogburn EL, Lunetta KL, *et al.* Demonstrating stratification in a European American population. *Nat Genet*. 2005; 37(8):868-72.
98. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994; 265(5181):2037-48.
99. Loh PR, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 2015; 47(3):284-90.
100. Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet*. 2013; 9(6):e1003566.
101. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Hidden 'risk' in polygenic scores: clinical use today could exacerbate health disparities. *bioRxiv*. 2018.
102. Bycroft C, Freeman C, Petkova D, *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018; 562(7726):203-209.
103. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*. 2017; 27(3S):S2-S8.
104. Gaziano JM, Concato J, Brophy M, *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016; 70:214-23.
105. Gravel S, Henn BM, Gutenkunst RN, *et al.* Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA*. 2011; 108(29):11983-8.
106. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012; 44(3):243-6.

107. Lawson DJ, Davies NM, Haworth S, *et al.* Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum Genet*. 2020; 139(1):23-41.
108. Ripatti S, Tikkanen E, Orho-Melander M, *et al.* A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet*. 2010; 376(9750):1393-400.
109. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, Ripatti S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013; 33(9):2261-6.
110. Abraham G, Havulinna AS, Bhalala OG, *et al.* Genomic prediction of coronary heart disease. *Eur Heart J*. 2016; 37(43):3267-3278.
111. Vilhjalmsson BJ, Yang J, Finucane HK, *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*. 2015; 97(4):576-92.
112. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med*. 2020; 26(4):549-557.
113. Khera AV, Chaffin M, Aragam KG, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018; 50(9):1219-1224.
114. Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*. 2019; 104(1):21-34.
115. Wood AR, Esko T, Yang J, *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014; 46(11):1173-86.
116. Berg JJ, Harpak A, Sinnott-Armstrong N, *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019; 8.
117. Sohail M, Maier RM, Ganna A, *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019; 8.
118. Robinson MR, Hemani G, Medina-Gomez C, *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat Genet*. 2015; 47(11):1357-62.
119. Haworth S, Mitchell R, Corbin L, *et al.* Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun*. 2019; 10(1):333.
120. Reisberg S, Iljasenko T, Lall K, Fischer K, Vilo J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One*. 2017; 12(7):e0179238.
121. Marnetto D, Parna K, Lall K, *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat Commun*. 2020; 11(1):1628.
122. Haggrén G, Halinen P, Lavento M, Raninen S, Wessman A. *Muinaisuutemme jäljet : Suomen esi- ja varhaishistoria kivikaudelta keskiajalle*. Helsinki: Gaudeamus; 2015.

123. Vahtola J. *Suomen historia : jääkaudesta Euroopan unioniin*. Helsinki: Otava; 2003.
124. Tallavaara M, Pesonen P, Oinonen M. Prehistoric population history in eastern Fennoscandia. *Journal of Archaeological Science*. 2010; 37(2):251-260.
125. Huurre M. *9000 vuotta Suomen esihistoriaa*. 5. ed. Helsinki: Otava; 1995.
126. Sjogren KG, Price TD, Kristiansen K. Diet and Mobility in the Corded Ware of Central Europe. *PLoS One*. 2016; 11(5):e0155083.
127. Paukkunen L. *Siirtokarjalaiset nyky-Suomessa*. Jyväskylän yliopiston yhteiskuntapolitiikan laitoksen tutkimuksia A,. Jyväskylän yliopisto; 1989.
128. Heikkilä E, Järvinen T. History and future lines of urbanization process in Finland. European Regional Science Association; 2002; Dortmund.
129. Population according to language 1980–2019. Population structure 2019. Statistics Finland. http://www.stat.fi/til/vaerak/2019/vaerak_2019_2020-03-24_tau_002_en.html. Accessed 6.5.2020.
130. Salminen T. Uralic Language Family. In: Moseley C, ed. *Encyclopedia of the World's Endangered Languages*. 2008 ed. New York: Routeledge; 2007.
131. Embleton S, Wheeletr ES. Finnish dialect atlas for quantitative studies. *Journal of Quantitative Linguistics*. 1997; 4(1-3):99-102.
132. Itkonen T. *Proto-Finnic final consonants : their history in the Finnic languages with particular reference to the Finnish dialects. I:1. The history of -k- in Finnish*. Helsinki: SUomalais-Ugrilainen Seura; 1965.
133. Itkonen T. *Nurmijärven murrekirja*. Suomalaisen kirjallisuuden seura; 1989.
134. Honkola T, Ruokolainen K, Syrjanen KJJ, *et al.* Evolution within a language: environmental differences contribute to divergence of dialect groups. *BMC Evol Biol*. 2018; 18(1):132.
135. Syrjänen K, Honkola T, Lehtinen J, Leino A, Vesakoski O. Applying Population Genetic Approaches within Languages. Research Article. *Language Dynamics and Change*. 2016; 6(2):235–283.
136. Lahermo P, Sajantila A, Sistonen P, *et al.* The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am J Hum Genet*. 1996; 58(6):1309-22.
137. Lahermo P, Savontaus ML, Sistonen P, *et al.* Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. *Eur J Hum Genet*. 1999; 7(4):447-58.
138. Sajantila A, Lahermo P, Anttinen T, *et al.* Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res*. 1995; 5(1):42-52.
139. Hannelius U, Salmela E, Lappalainen T, *et al.* Population substructure in Finland and Sweden revealed by the use of spatial coordinates and a small number of unlinked autosomal SNPs. *BMC Genet*. 2008; 9:54.
140. Salmela E, Lappalainen T, Fransson I, *et al.* Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS One*. 2008; 3(10):e3519.
141. Immigrants in the population. Statistics Finland. http://www.stat.fi/tup/maahanmuutto/maahanmuuttajat-vaestossa_en.html. Accessed 4.5.2020.

142. Cavalli-Sforza LL, Piazza A. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *Eur J Hum Genet*. 1993; 1(1):3-18.
143. Bauchet M, McEvoy B, Pearson LN, *et al.* Measuring European population stratification with microarray genotype data. *Am J Hum Genet*. 2007; 80(5):948-56.
144. Nelis M, Esko T, Magi R, *et al.* Genetic structure of Europeans: a view from the North-East. *PLoS One*. 2009; 4(5):e5472.
145. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491(7422):56-65.
146. Tambets K, Yunusbayev B, Hudjashov G, *et al.* Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol*. 2018; 19(1):139.
147. Nevanlinna HR. The Finnish population structure. A genetic and genealogical study. *Hereditas*. 1972; 71(2):195-236.
148. Workman PL, Mielke JH, Nevanlinna HR. The genetic structure of finland. *Am J Phys Anthropol*. 1976; 44(2):341-67.
149. Kittles RA, Perola M, Peltonen L, *et al.* Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet*. 1998; 62(5):1171-9.
150. Lappalainen T, Koivumaki S, Salmela E, *et al.* Regional differences among the Finns: a Y-chromosomal perspective. *Gene*. 2006; 376(2):207-15.
151. Raitio M, Lindroos K, Laukkanen M, *et al.* Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. *Genome Res*. 2001; 11(3):471-82.
152. Hedman M, Pimenoff V, Lukka M, Sistonen P, Sajantila A. Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. *Forensic Sci Int*. 2004; 142(1):37-43.
153. Palo JU, Hedman M, Ulmanen I, Lukka M, Sajantila A. High degree of Y-chromosomal divergence within Finland--forensic aspects. *Forensic Sci Int Genet*. 2007; 1(2):120-4.
154. Neuvonen AM, Putkonen M, Oversti S, *et al.* Vestiges of an Ancient Border in the Contemporary Genetic Diversity of North-Eastern Europe. *PLoS One*. 2015; 10(7):e0130331.
155. Lappalainen T, Laitinen V, Salmela E, *et al.* Migration waves to the Baltic Sea region. *Ann Hum Genet*. 2008; 72(Pt 3):337-48.
156. Kittles RA, Bergen AW, Urbanek M, *et al.* Autosomal, mitochondrial, and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am J Phys Anthropol*. 1999; 108(4):381-99.
157. Hedman M, Brandstatter A, Pimenoff V, *et al.* Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic Sci Int*. 2007; 172(2-3):171-8.
158. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. Genetic markers and population history: Finland revisited. *Eur J Hum Genet*. 2009; 17(10):1336-46.
159. Översti S, Majander K, Salmela E, *et al.* Human mitochondrial DNA lineages in Iron-Age Fennoscandia suggest incipient admixture and eastern

introduction of farming-related maternal ancestry. *Sci Rep*. 2019; 9(1):16883.

160. Jakkula E, Rehnstrom K, Varilo T, *et al.* The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet*. 2008; 83(6):787-94
161. Lappalainen T, Koivumaki S, Salmela E, *et al.* Regional differences among the finns: A Y-chromosomal perspective. *Gene*. 2006;376(2):207-215.
162. Lappalainen T, Hannelius U, Salmela E, *et al.* Population structure in contemporary Sweden--a Y-chromosomal and mitochondrial DNA analysis. *Ann Hum Genet*. 2009; 73(1):61-73.
163. Locke AE, Steinberg KM, Chiang CWK, *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*. 2019; 572(7769):323-328.
164. Fu Q, Li H, Moorjani P, *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. 2014; 514(7523):445-9.
165. Fu Q, Hajdinjak M, Moldovan OT, *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature*. 2015; 524(7564):216-9.
166. Gunther T, Malmstrom H, Svensson EM, *et al.* Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol*. 2018; 16(1):e2003703.
167. Lamnidis TC, Majander K, Jeong C, *et al.* Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat Commun*. 2018; 9(1):5018.
168. Mittnik A, Wang CC, Pfrengle S, *et al.* The genetic prehistory of the Baltic Sea region. *Nat Commun*. 2018; 9(1):442.
169. Ahola M, Salo KH, Mannermaa K. Almost Gone: Human Skeletal Material from Finnish Stone Age Earth Graves. A1 Journal article-refereed. *Fennoscandia Archaeologica*. 2016; (XXXIII):95-122.
170. Norio R. Finnish Disease Heritage I: characteristics, causes, background. *Hum Genet*. May 2003;112(5-6):441-56.
171. Norio R. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Hum Genet*. 2003; 112(5-6):457-69.
172. Lim ET, Wurtz P, Havulinna AS, *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet*. 2014; 10(7):e1004494.
173. Nikpay M, Goel A, Won HH, *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015; 47(10):1121-1130.
174. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014; 511(7510):421-7.
175. GBD 2016 Healthcare Access Quality Collaborators. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: a systematic analysis from the Global Burden of Disease Study 2016. *Lancet*. 2018; 391(10136):2236-2271.

176. Norio R. The Finnish Disease Heritage III: the individual diseases. *Hum Genet*. 2003; 112(5-6):470-526.
177. Koskinen S. THL's Morbidity Index. Statistical report 20/2019. THL's Morbidity Index 2014-2016. https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index. Accessed 2.6.2020.
178. Finnish institute for health and welfare. Sepelvaltimotauti-indeksi, ikävakioitu. http://www.terveytemme.fi/sairastavuusindeksi/2015/maakunnat_html/atlas.html?select=01&indicator=i0. Accessed 27.10.2018.
179. Puska P, Vartiainen E, Laatikainen T, Jousilahti P, Paavola M. *The Norther Karelia Project: From North Karelia To National Action*. National Institute for Health and Welfare (THL), in collaboration with the North Karelia Project Foundation; 2009.
180. Vartiainen E. The North Karelia Project: Cardiovascular disease prevention in Finland. *Glob Cardiol Sci Pract*. 2018; 2018(2):13.
181. Jussila A, Virta LJ, Salomaa V, Maki J, Jula A, Farkkila MA. High and increasing prevalence of inflammatory bowel disease in Finland with a clear North-South difference. *J Crohns Colitis*. 2013; 7(7):e256-62.
182. Borodulin K, Tolonen H, Jousilahti P, *et al.* Cohort Profile: The National FINRISK Study. *Int J Epidemiol*. 2018; 47(3):696-696i.
183. Goldstein JI, Crenshaw A, Carey J, *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics*. 2012; 28(19):2543-5.
184. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015; 4:7.
185. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. 2018. https://www.R-project.org/
186. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013; 10(1):5-6.
187. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014; 506(7488):376-81.
188. Liu JZ, van Sommeren S, Huang H, *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015; 47(9):979-986.
189. Shungin D, Winkler TW, Croteau-Chonka DC, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*. 2015; 518(7538):187-196.
190. Locke AE, Kahali B, Berndt SI, *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015; 518(7538):197-206.
191. Churchhouse C, Neale BM. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. 2017. http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank. Accessed 15.1.2018.

192. Athanasiadis G, Cheng JY, Vilhjalmsson BJ, *et al.* Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity. *Genetics*. 2016; 204(2):711-722.

193. Byrne RP, van Rheenen W, Project MinE ALS GWAS Consortium, van den Berg LH, Veldink JH, McLaughlin RL. Dutch population structure across space, time and GWAS design. *bioRxiv*. 2020.

194. Pankratov V, Montinaro F, Kushniarevich A, *et al.* Differences in local population history at the finest level: the case of the Estonian population. *Eur J Hum Genet*. 2020; 28(11):1580-1591.

195. Baharian S, Barakatt M, Gignoux CR, *et al.* The Great Migration and African-American Genomic Diversity. *PLoS Genet*. 2016; 12(5):e1006059.

196. Migration. Immigrants and integration. Statistics Finland. https://www.stat.fi/ tup/maahanmuutto/muuttoliike_en.html. Accessed 1.6.2020.

197. Lehtinen V, Joukamaa M, Lahtela K, *et al.* Prevalence of mental disorders among adults in Finland: basic results from the Mini Finland Health Survey. *Acta Psychiatrica Scandinavica*. 1990; 81(5):418-425.

198. Hovatta I, Terwilliger JD, Lichtermann D, *et al.* Schizophrenia in the genetic isolate of Finland. *Am J Med Genet*. 1997; 74(4):353-60.

199. Haukka J, Suvisaari J, Varilo T, Lonnqvist J. Regional variation in the incidence of schizophrenia in Finland: a study of birth cohorts born from 1950 to 1969. *Psychol Med*. 2001; 31(6):1045-53.

200. Perala J, Saarni SI, Ostamo A, *et al.* Geographic variation and sociodemographic characteristics of psychotic disorders in Finland. *Schizophr Res*. 2008; 106(2-3):337-47.

201. Kurki MI, Saarentaus E, Pietilainen O, *et al.* Contribution of rare and common variants to intellectual disability in a sub-isolate of Northern Finland. *Nat Commun*. 2019; 10(1):410.

202. Pietiläinen O. *Rare genomic deletio ns underlying schizophrenia and related neurodevelopmental disorders*. Helsinki: University of Helsinki; 2014.

203. Kaipiainen-Seppanen O, Aho K, Nikkarinen M. Regional differences in the incidence of rheumatoid arthritis in Finland in 1995. *Ann Rheum Dis*. 2001; 60(2):128-32.

204. Lehtinen P, Pasanen K, Kolho KL, Auvinen A. Incidence of Pediatric Inflammatory Bowel Disease in Finland: An Environmental Study. *J Pediatr Gastroenterol Nutr*. 2016; 63(1):65-70.

205. Regalado A. More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*. 2019.

206. Spector-Bagdady K, Fakih A, Krenz C, Marsh EE, Roberts JS. Genetic data partnerships: academic publications with privately owned or generated genetic data. *Genet Med*. 2019; 21(12):2827-2829.

207. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*. 2018;47(1):226-235.

208. Jorde LB, Wooding SP. Genetic variation, classification and 'race'. *Nat Genet*. 2004; 36(11 Suppl):S28-33.

209. Baker JL, Rotimi CN, Shriner D. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Sci Rep*. 2017; 7(1):1572.

210. Stra-tegic plan of the University of Helsinki 2021–2030: With the power of know-ledge – for the world. University of Helsinki. https://www.helsinki.fi/en/university/strategic-plan-2021-2030/strategic-plan-of-the-university-of-helsinki-2021-2030-with-the-power-of-knowledge-for-the-world. Accessed 14.5.2020.

211. Puttonen M. Suomalaiset jakaantuvat kymmeniin geneettisiin alaryhmiin - perimän erot noudattelevat murrerajoja, kertoo tutkimus. *Helsingin Sanomat*. 2017.

212. Waddington J. Uusi tutkimus löysi jopa 52 erilaista suomalaisten geeniryhmää - mihin sinä kuulut? *Ilta-Sanomat*. 2017.