ASSESSING FINNISH Y-CHROMOSOMAL HAPLOGROUPS USING GENOTYPING ARRAY DATA – TOWARDS UNDERSTANDING THE ROLE OF Y IN COMPLEX DISEASE

ANNINA PREUSSNER

MASTER'S THESIS

UNIVERSITY OF HELSINKI

FACULTY OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

GENETICS AND GENOMICS

MAY 2021

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Faculty of Biological and Environmental Sciences	Koulutusohjelma – Utbildingsprogram – Degree Programme Master's Programme in Genetics and Molecular Biosciences	
Tekijä – Författare – Author Annina Preussner		
Työn nimi – Arbetets titel – Title Assessing Finnish Y-chromosomal haplogroups using genotyping array data – Towards understanding the role of Y in complex disease		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetics and Genomics		
Työn laji – Arbetets art – Level Master's thesis	Aika – Datum – Month and year May 2021	Sivumäärä – Sidoantal – Number of pages 64

Tiivistelmä – Referat – Abstract

The Y chromosome has an essential role in the genetic sex determination in humans and other mammals. It contains a male-specific region (MSY) which escapes recombination and is inherited exclusively through the male line. The genetic variations inherited together on the MSY can be used in classifying Y chromosomes into haplogroups. Y-chromosomal haplogroups are highly informative of genetic ancestry, thus Y chromosomes have been widely used in tracing human population history. However, given the peculiar biology and analytical challenges specific to the Y chromosome, the chromosome is routinely excluded from genetic association studies. Consequently, potential impacts of Y-chromosomal variation on complex disease remain largely uncharacterized. Lately the access to large-scale biobank data has enabled to extend the Y-chromosomal genetic association studies. A recent UK Biobank study suggested links between Y-chromosomal haplogroup I1 and coronary artery disease (CAD) in the British population, but this result has not been validated in other datasets. Since Finland harbours a notable frequency of Y-chromosomal haplogroup I1, the relationship between haplogroup I1 and CAD can further be inferred in the Finnish population using data from the FinnGen project.

The first aim of this thesis was to determine the prevalence of Y-chromosomal haplogroups in Finland and characterize their geographical distributions using genotyping array data from the FinnGen project. The second aim was to assess the role between Finnish Y-chromosomal haplogroups and coronary artery disease (CAD) by logistic regression.

This thesis characterized the Y-chromosomal haplogroups in Finland for 24 160 males and evaluated the association between Y-chromosomal haplogroups and CAD in Finland. The dataset used in this study was extensive, providing an opportunity to study the Y-chromosomal variation geographically in Finland and its role in complex disease more accurately compared to previous studies. The geographical distribution of the Y-chromosomal haplogroups was characterized on 20 birth regions, and between eastern and western areas of Finland. Consistent with previous studies, the results demonstrated that two major Finnish Y-chromosomal haplogroup lineages, N1c1 and I1, displayed differing distributions within regions, especially between eastern and western Finland. Results from logistic regression analysis between CAD and Y-chromosomal haplogroups suggested no significant association between haplogroup I1 and CAD. Instead, the major Finnish Y-chromosomal haplogroup N1c1 displayed a decreased risk for CAD in the association analysis when compared against other haplogroups. Moreover, this thesis also demonstrated that the association results were not straightforwardly comparable between populations. For instance, haplogroup I1 displayed a decreased risk for CAD in the FinnGen dataset when compared against haplogroup R1b, whereas the same association was reported as risk increasing for CAD in the UK Biobank.

Overall, this thesis demonstrates the possibility to study the genetics of Y chromosome using data from the FinnGen project, and highlights the value of including this part of the genome in the future complex disease studies.

Avainsanat – Nyckelord – Keywords
Y chromosome, Y-chromosomal haplogroups, FinnGen, coronary artery disease, complex disease, logistic regression

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Taru Tukiainen & Jaakko Leinonen

Säilytyspaikka – Förvaringställe – Where deposited
HELDA – Digital Repository of the University of Helsinki

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta – Fakultet – Faculty Bio- ja ympäristötieteellinen Tiedekunta	Koulutusohjelma – Utbildingsprogram – Degree Programme Genetiikan ja molekulaaristen biotieteiden maisteriohjelma	
Tekijä – Författare – Author Annina Preussner		
Työn nimi – Arbetets titel – Title Suomalaisten Y-kromosomaalisten haploryhmien määrittäminen genotyyppisirudatasta – Kohti ymmärrystä Y-kromosomin roolista monitekijäisissä sairauksissa		
Oppiaine/Opintosuunta – Läroämne/Studieinriktning – Subject/Study track Genetiikka ja Genomiikka		
Työn laji – Arbetets art – Level Pro gradu	Aika – Datum – Month and year Toukokuu 2021	Sivumäärä – Sidoantal – Number of pages 64

Tiivistelmä – Referat – Abstract

Y-kromosomilla on olennainen rooli ihmisen ja muiden nisäkkäiden geneettisessä sukupuolenmäärityksessä. Se sisältää miehille ominaisen MSY-alueen, joka ei rekombinoi, ja siten periytyy yksinomaan miesten sukulinjan kautta. MSY-alueen geneettiset variaatiot periytyvät yhdessä ja siten Y-kromosomit voidaan lajitella haploryhmiin. Y-kromosomaaliset haploryhmät sisältävät informaatiota geneettisestä alkuperästä, joten Y-kromosomeja on käytetty laajalti ihmisen historian jäljittämisessä. Y-kromosomille ominaiset biologiset ja analyyttiset haasteet huomioon ottaen Y-kromosomi jätetään usein rutiininomaisesti pois geneettisistä assosiaatiotutkimuksista. Näin ollen Y-kromosomaalisen geneettisen vaihtelun mahdolliset vaikutukset monitekijäisissä sairauksissa ovat suurelta osin määritelmättä. Viime aikoina Y-kromosomin geneettisten assosiaatiotutkimusten laajentaminen on ollut mahdollista laajamittaisten biopankkidatojen avulla. Äskettäisessä Ison-Britannian biopankkitutkimuksessa ehdotettiin assosiaatiota Y-kromosomaalisen I1-haploryhmän ja sepelvaltimotaudin (CAD) välillä Ison-Britannian väestössä, mutta tätä tulosta ei ole vahvistettu muissa aineistoissa. Koska I1-haploryhmä on yleinen Suomen väestössä, voidaan tätä I1-haploryhmän ja sepelvaltimotaudin välistä yhteyttä tutkia FinnGen-projektin datan avulla.

Tämän tutkielman ensimmäisenä tavoitteena oli selvittää Y-kromosomaalisten haploryhmien yleisyys Suomessa, ja luonnehtia haploryhmien maantieteellistä jakautumista Suomessa käyttämällä FinnGen-projektin genotyyppisirudataa. Toisena tavoitteena oli tutkia suomalaisten Y-kromosomaalisten haploryhmien ja sepelvaltimotaudin (CAD) välistä yhteyttä logistisella regressioanalyysillä.

Tämä tutkielma määritteli Y-kromosomaalisia haploryhmiä Suomessa 24 160 miehelle ja tutki näiden Y-kromosomaalisten haploryhmien ja sepelvaltimotaudin välistä yhteyttä. Tutkimuksen aineisto oli moninkertainen aiempiin tutkimuksiin verrattuna, tarjoten mahdollisuuden tutkia Y-kromosomaalisen variaation maantieteellistä jakaumaa Suomessa ja tämän merkitystä monitekijäisille taudeille huomattavasti aiempia tutkimuksia tarkemmin. Y-kromosomaalisten haploryhmien maantieteellinen jakauma karakterisoitiin sekä 20 syntymäalueella että Suomen itä- ja länsiosien välillä. Aikaisempien tutkimusten mukaisesti tulokset osoittivat, että kahden yleisen suomalaisen Y-kromosomaalisten N1c1- ja I1-haploryhmien välillä ilmeni eroja esiintyvyydessä alueellisesti, etenkin Itä- ja Länsi-Suomen välillä. Tulokset logistisista regressioanalyyseistä sepelvaltimotaudin ja Y-kromosomaalisten haploryhmien välillä eivät osoittaneet merkitsevää yhteyttä I1 haploryhmän ja sepelvaltimoataudin välillä. Sen sijaan yleisimmällä suomalaisella Y-kromosomaalisella N1c1-haploryhmällä oli vähentynyt riski sepelvaltimotaudille assosiaatioanalyysissä verrattuna muihin haploryhmiin. Tulokset osoittivat myös, että Y-kromosomaalisten haploryhmien assosiaatiotulokset eivät olleet suoraan vertailukelpoisia populaatioiden välillä. Esimerkiksi sepelvaltimotaudin ja Y-kromosomaalisen I1-haploryhmän välinen assosiaatio osoitti alentuneen sepelvaltimotautiriskin verrattaessa R1b-haploryhmään FinnGen aineistossa, kun taas saman assosiaation ilmoitettiin lisäävän sepelvaltimotauti-riskiä Isossa-Britanniassa.

Kaiken kaikkiaan tämä tutkielma osoittaa mahdollisuuden tutkia Y-kromosomien genetiikkaa FinnGen-projektin datan avulla ja korostaa tämän genomin osan sisällyttämisen arvoa tulevissa monitekijäisten sairauksien tutkimuksissa.

Avainsanat – Nyckelord – Keywords
Y-kromosomi, Y-kromosomaaliset haploryhmät, FinnGen, sepelvaltimotauti, monitekijäiset sairaudet, logistinen regressio

Ohjaaja tai ohjaajat – Handledare – Supervisor or supervisors
Taru Tukiainen & Jaakko Leinonen

Säilytyspaikka – Förvaringställe – Where deposited
HELDA – Helsingin yliopiston digitaalinen arkisto

Muita tietoja – Övriga uppgifter – Additional information

CONTENTS

ABBREVIATIONS

ATC	Anatomical Therapeutic Chemical
AZF	Azoospermia factor
BMI	Body mass index
BP	Before present
CAD	Coronary artery disease
CF	Central Finland
CI	Confidence interval
CKA	Ceded Karelia
CO	Central Ostrobothnia
FDH	Finnish disease heritage
FIMM	Institute for molecular medicine Finland
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
ISOGG	International society of genetic genealogy
KA	Kainuu
KH	Kanta-Häme
KY	Kymenlaakso
LA	Lapland
LOY	Loss of Y chromosome
Mb	Million base pairs
MSY	Male-specific region of the Y chromosome
N	Sample size
NE	Northeast
NK	North Karelia
NO	North Ostrobothnia
NS	North Savonia
OR	Odds ratio
OS	Ostrobothnia

OST	Central Ostrobothnia, Ostrobothnia and South Ostrobothnia
PAR	Pseudoautosomal region
PC	Principal Component
PCA	Principal Component analysis
PH	Päijät-Häme
PI	Pirkanmaa
PRS	Polygenic risk score
SA	Satakunta
SK	South Karelia
SKA	South Karelia and Kymenlaakso
SNP	Single nucleotide polymorphism
SO	South Ostrobothnia
SRY	Sex-determining region Y
SS	South Savonia
SW	Southwest
SWF	Southwest Finland
TAV	Tavastia
UKB	UK Biobank
UU	Uusimaa
ÅLA	Åland

1 INTRODUCTION

1.1 THE HUMAN Y CHROMOSOME

1.1.1 Biological role

The human genome consist of DNA, most of which is located at the cell nucleus, with the exception of a small amount of non-nuclear DNA located in the mitochondria. The nuclear DNA is organized into 46 chromosomes, consisting of 22 autosomal chromosome pairs and one pair of sex chromosomes. Apart from the autosomal genome, the sex chromosomes X and Y are differently distributed between males and females. The male genome contains both X and Y chromosomes (46, XY), while the female genome contains two X chromosomes (46, XX). Biologically the most important Y-linked gene is the *SRY* (sex-determining region Y), which is responsible for sex differentiation during embryonic development. The presence of *SRY* in males initiates the gonads to develop as testis, whereas the lack of *SRY* gene expression results in the development of ovaries (Kashimada and Koopman, 2010).

1.1.2 Structure of the Y chromosome

The human X and Y chromosome have evolved from an ordinary pair of autosomes, which trough millions of years of evolution gained genetic and morphological differences (Bachtrog, 2006). The human Y chromosome is one of the smallest chromosomes having a length of 57 Mb, whereas its chromosome pair X is approximately three times the size of the Y, 156 Mb (Howe et al., 2020). In addition, the Y chromosome differs remarkably by its structure compared to the rest of the genome. The Y chromosome contains a male-specific region (MSY), which is unique to the Y chromosome and does not undergo recombination (Skaletsky et al., 2003). Meiotic recombination is the exchange of genetic material between chromosome pairs in the formation of gametes (Baudat et al., 2013). In every generation, the genetic material is broken down by recombination between homologous chromosomes and is randomly transmitted to the next generation, leading to novel genetic combinations of the DNA in the offspring. However, the MSY does not undergo recombination and is inherited through the paternal line as a haploid segment unlike the rest of the chromosomes (Figure 1). Nevertheless, in addition to 54 Mb of MSY, the Y chromosome contains 3 Mb of homologous DNA shared with the X chromosome. These pseudoautosomal regions, PAR1 and PAR2,

are located at telomeric regions of the X and Y chromosomes, short and long arm respectively (Helena Mangs and Morris, 2007; Skaletsky et al., 2003). The shared homology, especially on pseudoautosomal region 1, is essential for correct sex chromosome segregation in male meiosis (Helena Mangs and Morris, 2007).

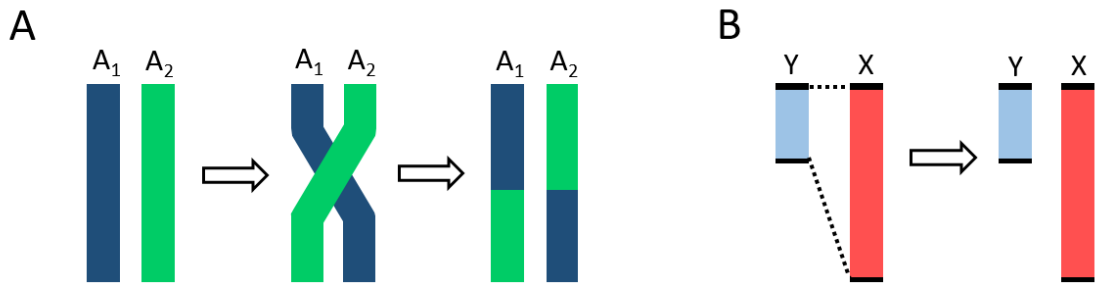

Figure 1. A) Recombination between chromosomes A1 and A2 results in new combinations of the genetic material. Chromosomes A1 and A2 represent an autosomal chromosome pair, or correspondingly a pair of X chromosomes. B) Recombination does not occur between Y and X chromosomes, with the exception of pseudoautosomal regions (black) on the both ends of the sex chromosomes. The light blue region illustrates the male-specific region of the Y (MSY), which is inherited through the paternal line a continuous DNA segment.

1.1.3 Genetic content

Since the Y chromosome contains large quantities of repetitive and heterochromatic sequences, it essentially contains a small amount of active genes (Skaletsky et al., 2003). Heterochromatin is a tightly packed form of DNA in genetically inactive regions. In contrast, euchromatin is a more lightly packed form of DNA found in active regions of the genome. Heterochromatin is typically found at centromeres and telomeric regions of chromosomes, but in addition to these, the Y chromosome contains a long heterochromatic block that comprises the bulk of the Y-chromosome long arm (Skaletsky et al., 2003). Heterochromatin covers the majority of the MSY sequence, approximately 31 Mb, whereas only 23 Mb consists of euchromatin (Howe et al., 2020; Skaletsky et al., 2003). From the total of 565 genes on the Y chromosome, only 65 are annotated as protein-coding genes, whereas the rest consists of non-coding genes and pseudogenes (Howe et al., 2020). Several of the 65 protein-coding genes belong to the same protein families, leaving only 27 genes coding for distinct MSY proteins (Howe et al., 2020). Most of the MSY genes are expressed in the in the reproductive tissue, and are essential for male sex development and spermatogenesis, whereas some genes are also expressed ubiquitously in non-reproductive tissues (Skaletsky et al., 2003). Approximately half of the protein-coding genes or gene-families on the MSY genes display their

highest gene expression profiles in the testis, whereas half of them have their highest expression in non-reproductive tissues (Godfrey et al., 2020). There are 17 genes on the Y chromosome, which share an ancestral (non-identical) homolog with the X chromosome. These genes are highly dosage sensitive, controlling many essential cell functions, such as transcription, translation and protein stability (Bellott et al., 2014).

1.1.4 Y-chromosomal haplogroups

Although two humans share on average 99.9% of their genome, there are around four to five million genomic variants to be found between individuals (1000 Genomes et al., 2015). Genetic variants are variations within the DNA, and they can be classified according to their size and type of the variation. Small-scale variations include single nucleotide variants, short tandem repeats (also known as microsatellites) and small insert-deletions, whereas larger variations include copy-number, structural, and chromosomal variations (Sharp et al., 2006). Approximately 96% of genetic variations in humans consist of single nucleotide polymorphisms (SNPs) (1000 Genomes et al., 2015), which are variations within one base of the DNA (A, T, G or C). The distinction between a variant and a polymorphism is that classically polymorphism is defined as having a frequency of at least 1% in the population (Sharp et al., 2006).

Genetic variations that are not separated by genetic recombination, but inherited together from one parent to offspring, construct a haplotype. Since the paternally inherited MSY escapes recombination, it can uniquely be used in classifying Y chromosomes into haplotypes (Jobling and Tyler-Smith, 2003). Individuals that share identical patterns of genetic variations on the MSY, belong to the same Y-chromosomal haplotype. If two haplotypes share a common ancestor in the paternal line, they can be classified into the same haplogroup. Generally, all Y-chromosomal haplotypes present in modern populations share a common ancestor, haplogroup A (Y-chromosomal Adam). This ancestral haplogroup has been suggested to have evolved in Africa around 120 000 years ago during the modern human (*Homo sapiens*) history (Hallast et al., 2015). Distinct Y-chromosomal haplotypes have established at different time points during human history through the accumulation of DNA mutations on the MSY. Especially the genetic variations emerged during the past 50 000 years after the out-of-Africa dispersal of humans (Poznik, 2016), are informative for tracing human population history and migration patterns of males (Lippold et al., 2014; Underhill et

al., 2001). In modern populations, the prevalence of existing Y-chromosomal haplogroups depends highly on the geographical location (Jobling and Tyler-Smith, 2003).

Y-chromosomal haplotypes can be defined by sequencing the Y-chromosomal DNA or by genotyping a known set of genetic variations, the latter method being more cost-effective in large datasets. Genotyping is performed on microarrays, which contain information of the genetic positions of interest, and allow detection of the allelic state of each marker position in the sample DNA. The markers used in genotyping can be binary SNP markers or microsatellites, but since SNPs are more common in the human genome, they are usually preferred as markers in genotyping (Kim, S. and Misra, 2007).

By estimating mutation rates for different haplogroup-defining variants, the Y-chromosomal haplotypes can be arranged into a phylogeny presenting the relationships and estimated emergence points for each Y-chromosomal haplogroup (Figure 2) (Hallast et al., 2015; Y Chromosome Consortium, 2002). The Y-chromosomal haplogroups follow a unified nomenclature system described by the Y Chromosome Consortium, each haplogroup containing information of its phylogenetic position (Y Chromosome Consortium, 2002). The first capital letter (from A to T) defines the major haplogroups, and additional lower case letters and numbers define their subhaplogroups (e.g. I1a). The subhaplogroups differ from their ancestral haplogroup by subsequent mutations (Y Chromosome Consortium, 2002). The current phylogenetic tree of human MSY phylogeny encompasses well-defined haplogroups, and it is continuously refined by the detection of new Y-chromosomal polymorphisms (Hallast et al., 2015; Karafet et al., 2008; Poznik, 2016).

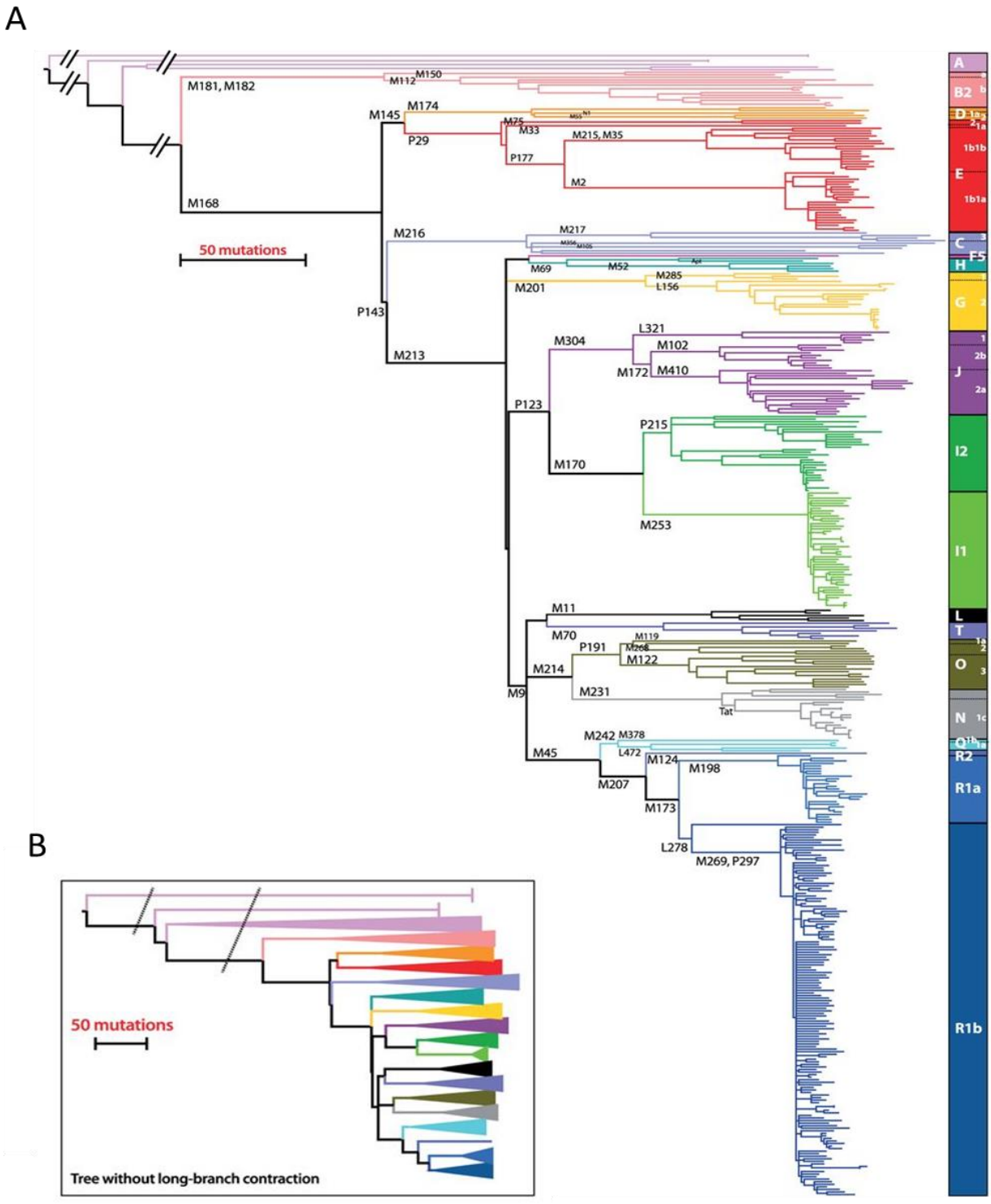

Figure 2. A) Phylogeny of Y-chromosomal haplogroups. Major Y-chromosomal haplogroup lineages are indicated by colors, and haplogroup-defining mutations are indicated on the branches. B) Simplified phylogenetic tree showing true lengths of deep-rooting branches. Figure adapted from Hallast et al. (2015) under the Creative Commons Attribution 4.0 International License.

1.2 THE ROLE OF Y CHROMOSOME IN DISEASE AND METHODS TO STUDY GENETIC ASSOCIATIONS

1.2.1 The role of Y chromosome in disease

Most of the Y chromosome genes have an important role in maintaining male fertility, therefore the dysfunction of these genes usually have severe consequences in males. The deletion or translocation of the sex-determining gene *SRY* causes severe disorders in sex development and results in failure of gonadal development (Jobling and Tyler-Smith, 2017). Moreover, microdeletions in other genes essential for spermatogenesis are found to cause infertility. Microdeletions may for instance, occur though gene conversion in the long arm of the Y chromosome, which contains many repeated ampliconic and palindromic sequences. Microdeletions in the AZF-locus (Azoospermia factor) located at the long arm of the Y chromosome, is one of the major causes of male infertility (Colaco and Modi, 2018; Jobling and Tyler-Smith, 2017). Moreover, the Y chromosome can be lost during mitotic cell division, resulting in mosaic loss of the Y chromosome (LOY). Mosaic LOY is a pathogenic condition, and it has been linked to various disorders, such as cancers and cardiovascular diseases (Barros et al., 2020). However, these conditions described above are rare and they are not transferred to the next generation. Less is known for instance, about how common Y-chromosomal genetic variations affect the development of complex disease, which are caused by the interaction of multiple genes and environmental factors.

Many common diseases display sex differences at some level in their prevalence, progression, age-of onset or severity (Khramtsova et al., 2019; Ober et al., 2008). Various diseases, such as cancers, neurological, and cardiovascular diseases exhibit sex differences. For instance, Parkinson disease and cardiovascular disease are more common in males, whereas Alzheimer disease is more common in females (Ober et al., 2008). Sex differences in complex diseases are established by complex interactions of genetic, hormonal and environmental factors (Khramtsova et al., 2019).

Although the role of Y chromosome in complex disease is not well understood, some studies have evaluated the link between Y-chromosomal genetic variation and complex disease. These association studies usually have considered Y-chromosomal haplogroups, rather than single isolated variants on the MSY. Studies have considered for instance the role of Y-chromosomal haplogroups in cardiovascular disease (Eales et al., 2019; Voskarides et al., 2014), behavioral traits (Howe et al., 2017), prostate cancer (Patel et al., 2018), and viral infections (Sezgin et al., 2009). Few of these

reported associations, including cardiovascular disease (Eales et al., 2019) and faster human immunodeficiency virus (HIV) progression (Sezgin et al., 2009), have been linked to a common European Y-chromosomal haplogroup lineage I. These associations for haplogroup I have been suggested to arise from altered immune responses due to differences in inflammation and immune-related gene expression in macrophages (Eales et al., 2019). However, the underlying biological mechanisms and causal variants are not yet well characterized for the findings. Overall, the associations between Y-chromosomal haplogroups and complex disease have been difficult to confirm due to several challenges discussed in more detail in section 1.2.3.

1.2.2 Methods to study genetic associations

Genetic association studies are utilized in identifying disease-associated genetic loci by testing correlation between genetic variation and disease status (Hirschhorn et al., 2002). SNPs are commonly used markers in genetic association testing, but also other genetic variations can be used. Genome-wide association studies (GWAS) are a broader approach to identify genetic risk factors for complex disease through the whole genome by statistically testing the effects of up to millions of genetic variants separately to the disease or trait of interest (Tam et al., 2019). Currently GWAS have proven successful in providing information of the genetic architecture underlying a wide range of human complex disease (Buniello et al., 2019).

The most commonly used approach in a genetic association studies is a retrospective study setting, in which unrelated individuals are divided into cases and controls by their phenotype. The allele frequencies of the genetic variants are then statistically compared between cases and controls to identify disease-associated genetic differences (Hirschhorn et al., 2002) (Figure 3). Genetic associations are typically calculated under an additive model, where the phenotype Y is assumed to change linearly by the allele copies in a genetic locus X.

$$Y \sim \mu + X\beta + \varepsilon \quad (1)$$

where μ is the mean genotype, ε the error term and β the effect estimate of the genotype X. When measuring binary phenotypes, the genetic effect β is traditionally expressed as odds ratio OR = e^{β}.

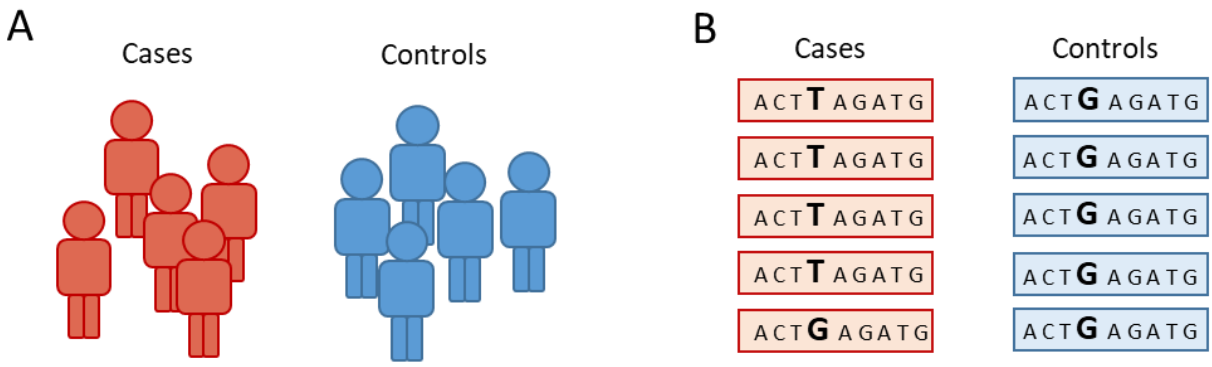

Figure 3. The idea of a case-control study setting. A) Samples are divided into cases (red) and controls (blue) by their phenotype. B) The differences between cases and controls are assessed on genotype level. Genetic variations observed with a higher frequency in cases suggests that the allele or genotype is associated with the phenotype of interest.

The statistical significance of an association is evaluated by a p-value. The observations that reach a p-value threshold and show an inconsistency with the null hypothesis can be labelled as statistically significant. In genetic association studies, the null hypothesis states that there are no effects between the genotype and phenotype of interest, whereas the alternative hypothesis states that the genotype is associated with the phenotype of interest.

$$H_0 : \beta = 0 \qquad (2)$$

$$H_1 : \beta \neq 0$$

Often a nominal p-value threshold of 0.05 is applied for significance testing. However, when performing multiple tests for the null hypothesis, e.g. when testing multiple SNPs for their association, the likelihood of observing a false discovery increases. A typical method to control for multiple testing is Bonferroni correction, where the p-value threshold is adjusted by the number of independent tests (Armstrong, 2014). For instance, a generally accepted p-value threshold in GWAS is $0.05/1\,000\,000 = 5 \times 10^{-8}$, since there are approximately one million independent genetic variants along the genome (Tam et al., 2019). A significant genetic association may be interpreted as either as direct association, indirect association or a false positive result. Direct association implies the genetic locus identified is a causal variant for the disease, whereas an indirect association is caused if the identified locus is in linkage disequilibrium with the true variant. Instead, a false positive result refers to an identified locus that is not responsible for the variation in the trait. False positives may arise by chance or by systematic confounding (Lewis and Knight, 2012).

Since complex disorders are influenced by both, genetic and environmental factors, the interpretation of genetic association studies of complex disorders requires consideration of many potential confounders that might affect the disease phenotype. Essentially, genetic association analyses should be stratified according to possible confounders, by including them as covariates in the regression model

$$Y \sim \mu + Z\gamma + X\beta \quad (3)$$

where Z is a confounder of the association between genotype X and phenotype Y.

A typical confounder in GWAS is population structure, which may arise because of differences in genetic ancestries between cases and controls (Price et al., 2006). If samples from different ancestries are compared, the differences in genetic ancestry might partly correlate with the phenotype of interest, thereby resulting in observing a false positive association. Nevertheless, population structure can be adjusted in genetic association testing. Population structure is usually adjusted by autosomal principal components (PCs) that are produced by principal component analysis (PCA) of genetic markers. PCA reduces the dimensionality of large datasets, while capturing as much variation as possible into the principal components. Thus, these principal components can be used in explaining heterogeneity observed between samples, that may for instance, result from population structure (Price et al., 2006).

1.2.3 Challenges in Y-chromosomal genetic association studies

Although genome-wide association studies aim to identify genetic risk factors across the whole genome, the Y chromosome is commonly excluded from these analyses due to several reasons. The genetic content of the Y chromosome is low, and most of the genes have been linked to male-specific features, such as spermatogenesis. Since the Y chromosome is only carried by males, this leads to reduced statistical power compared to other chromosomes in genetic association studies, because only approximately half of the samples can be included in the analysis. Moreover, the Y chromosome differs from the rest of the genome also in terms of physical structure, which creates additional challenges for its analysis. Since the Y chromosome is haploid, for instance, variant filtering pipelines used for the autosomal genome are not directly suitable for the Y-chromosomal variants, requiring separate analysis from the rest of the genome. In addition, the special features in Y chromosome recombination pose challenges for example to imputation, referring to statistically inferring genotypes that are not directly genotyped (Marchini and Howie, 2010). Due to these

limitations and challenges, the convention in the field has thus been to exclude the Y chromosome from GWAS.

Although there are some reports suggesting links between Y-chromosomal genetic variation and complex disease, the interpretation of the few genetic association studies focusing on the Y chromosome remain challenging. One concern rises especially from the geographical clustering of Y-chromosomal variants, which may complicate association studies focusing on Y-chromosomal haplogroups (Erzurumluoglu et al., 2018). The geographical clustering of Y-chromosomal haplogroups limits the possibility to study the suggested associations in other populations. For instance, the suggested link between Y-chromosomal haplogroup O3 and prostate cancer (Paracchini et al., 2003) may be studied only in East Asian populations, since the haplogroup O3 in extremely rare in other populations (Shi et al., 2005).

Moreover, the geographical clustering of Y-chromosomal haplogroups creates challenges in interpretation of the association results. Diverse populations do not only carry certain Y-chromosomal variations, some of which may be population specific, but also harbor population specific autosomal genetic profiles, which may correlate with the Y-chromosomal variants. Thereby it is challenging to distinguish the real effects of the Y-chromosomal genetic variants from the rest of the genome, and the association studies might not replicate in other populations, making validation of the results challenging. For instance, Paracchini et al. (2003) reported haplogroup O3 to associate with a higher risk to prostate cancer in Japanese males, however, such association for the same haplogroup was not found in Korean males (Kim, W. et al., 2007). Finally, identifying causal variants on the Y chromosome remains challenging, since all the MSY variants segregating with a given haplotype are in linkage disequilibrium. Thereby, further identification of the exact causal variants on the Y chromosome ultimately requires careful functional analyses (Parker et al., 2020).

1.3 HETEROGENEITY OF THE FINNISH POPULATION

1.3.1 Genetic characteristics of the Finnish population

The Finnish population has been a target for several population and medical genetics studies. When considering Europe, the northern populations display higher genetic linkage-disequilibrium patterns and lower genetic diversity compared to those of southern Europe (Lao et al., 2008; Salmela et al.,

2008). Especially the Finnish population can be considered a genetic isolate when comparing the autosomal genome to other European populations (Lao et al., 2008; Salmela et al., 2008). Founder effects of small numbers of early settlers, and genetic drift have shaped the genetic structure of the Finnish population. The Finnish population is enriched in many genetic variants that are seen at a lower frequency elsewhere in the world, and alternatively lacks some genetic variants that are more common elsewhere (1000 Genomes et al., 2015; Lim et al., 2014).

Although the Finnish population is considered a genetic isolate, there exists abundant genetic sub-structure within the Finnish population (Kerminen et al., 2017). After the early settlers arrived to the area of modern day Finland after the last glacial period approximately 10 500 years ago, the land has faced several cultural and geographical changes, including for example the arrival of combed ware and corded ware cultures and continuous land uplift (Cramp et al., 2014; Tambets et al., 2018). During the past centuries, Finland has resided at the border between southwestern populations of Scandinavia and Baltics and eastern populations of Russia, in addition to the indigenous Sami people in the north (Kerminen et al., 2017). Especially the genetic influences from southwestern and eastern directions can be observed in the population structure of present day Finland, as the main genetic division in Finland can be seen between eastern and western regions. Over the last millennium, before gaining independence in 1917, Finland was under the rule of either Sweden or Novgorod (present day Russia), and interestingly the genetic division of Finns follows a former political borderline between Sweden and Novgorod set in the Treaty of Nöteborg in 1323 (Kerminen et al., 2017). According to some estimates, the genetic difference between eastern and western Finns is stronger than for instance, the difference observed between British and Germans (Salmela et al., 2008).

1.3.2 Y-chromosomal haplogroups in Finland

Finland is enriched in two Y-chromosomal haplogroups, N1c1 (N-M178) and I1 (I-M170), which correspondingly represent dissimilar distributions between eastern and western Finland (Lappalainen et al., 2006; Neuvonen et al., 2015). Kittles et al. (1998) first characterized the divergence of Finnish Y-chromosomal haplotypes in a study of 280 males from nine Finnish provinces. Further studies by Lappalainen et al. (2006) and Neuvonen et al. (2015) extended these findings, respectively including 536 and 584 males in their research. The Y-chromosomal haplogroups N1c1 and I1 are described to make up around 90% of all Finnish Y chromosomes

(Lappalainen et al., 2006; Neuvonen et al., 2015). The Y-chromosomal haplogroup N1c1 has been described to be more frequent in northeastern Finland, whereas I1 is enriched in the southwestern Finland. According to the study of Lappalainen et al. (2006), the frequencies of N1c1 and I1 are 70.9% and 19.6% in northeastern Finland, but 41.3% and 41.3% in southwestern Finland, respectively. Less frequently seen haplogroups in Finland comprise haplogroups R1a, R1b and I2, whereas for example haplogroups E1b1, Q and J seem to be extremely rare in Finland (Lappalainen et al., 2006; Neuvonen et al., 2015).

The two major Finnish haplogroup lineages, N and I, are estimated to have evolved approximately 20 000 BP (before present) and 20 600 BP, respectively (Hallast et al., 2015). These Y-chromosomal lineages have different geographical origins, haplogroup N having its suggested origin in Southeast Asia (Rootsi et al., 2007), and haplogroup I in Europe (Rootsi et al., 2004). Haplogroup N1c1 (N-M178) is geographically most frequent in Finland and it has an estimated origin around 4 600 BP (Hallast et al., 2015) in Southern Siberia (Rootsi et al., 2007). Haplogroup I1 (N-M253) is geographically most enriched in Scandinavia, and its origin is dated approximately 3 500 BP (Hallast et al., 2015; Rootsi et al., 2004).

1.3.3 East-West health disparities in Finland

The Finnish population shows geographic variation in disease prevalence, many diseases having a higher prevalence in eastern Finland. Several rare monogenic diseases, classified as the Finnish disease heritage (FDH), show a higher prevalence especially in eastern Finland. The FDH comprises 36, mostly autosomal recessive disorders that are enriched in the Finnish population (Norio, 2003). However, the FDH diseases are relatively rare in the overall population. Nevertheless, many common complex diseases show also disparities between eastern and western Finland (https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index 1.4. 2021) (Figure 4). For instance, coronary artery disease, musculoskeletal disorders and mental health-related conditions have a higher morbidity index in the eastern Finland, especially in northern Savonia (https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index 1.4.2021). However, not all conditions are enriched exclusively to eastern Finland, and for instance, cancers display a high incidence in southern Finland (https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index, 1.4.2021). These geographic health differences are influenced by many genetic, lifestyle, and environmental factors and the full

understanding of the mechanisms behind these geographical health disparities are not fully characterized.

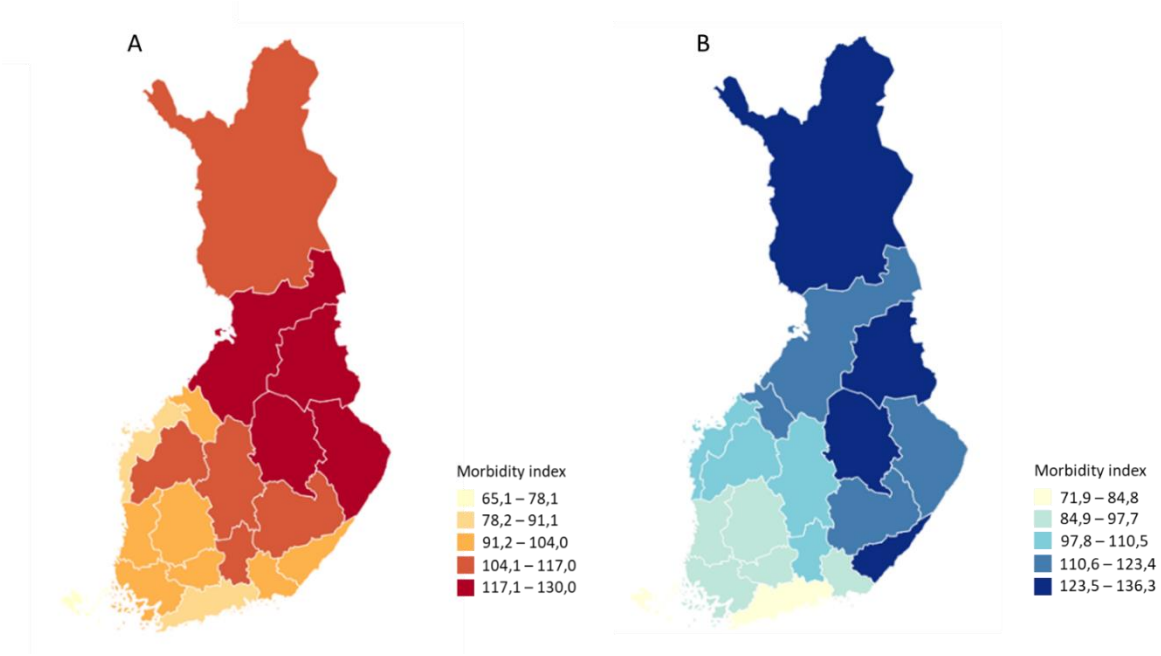

Figure 4. A) Age-adjusted morbidity index combined for seven disease groups in 2014 – 2016: cancers, coronary heart disease, cerebrovascular disease, musculoskeletal disorders, mental disorders, injuries and dementia. B) Age-adjusted morbidity index for coronary artery disease. Incidence rate of 100 describes the mean morbidity in Finland. Figures were generated on THL web resource http://www.terveytemme.fi/sairastavuusindeksi/2016/maakunnat_html_profiili/atlas.html?select=01&indicator=i0 (1.4.2021).

1.3.3.1 Coronary artery disease

In Finland, coronary artery disease (CAD) displays a higher incidence in eastern Finland compared to western Finland (Figure 4B). Although the mortality for CAD has decreased enormously through the past 50 years (Vartiainen, 2018), diseases of the circulatory system remain one of the leading causes of mortality in Finland (https://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_001_en.html 1.4.2021). In addition to the geographical differences of CAD mortality in Finland, the mortality is higher in males compared to females. For instance, in 2019 the rates for ischemic heart disease related deaths in Finland were 1825 per 10 000 deaths in males (https://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_002_en.html 1.4.2021), but 1372 per 10 000 deaths in females (https://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_003_en.html 1.4.2021).

The major cause of CAD is coronary atherosclerosis, which is a condition where plaque accumulates on the inner walls of blood vessels of the heart, thus reducing blood flow to the heart muscle. Major complications of CAD include angina, myocardial infarction, cardiac arrhythmias and heart failure (Kessler et al., 2016). Known risk factors for CAD include for instance, hypertension, hyperlipidemia, smoking, obesity and diabetes mellitus (Malakar et al., 2019).

Large-scale GWAS have identified more than 60 genomic loci associated with CAD, which however, collectively explain only less than 20% of the heritability of CAD (Hartiala et al., 2017). The strongest association for CAD ($p = 2.3\times10^{-98}$, OR = 1.21) is located to chromosome 9p21 (lead SNP rs2891168) (Hartiala et al., 2017). However, most of the CAD associated loci are located in non-coding regions and have modest effect sizes (OR < 1.1) (Hartiala et al., 2017). The observation that CAD has overall small heritability as estimated from GWAS results, suggests either the presence of unobserved genetic variants with small effect sizes, rare variants, gene-environment interactions, or other unknown biological mechanisms that GWAS are unable to capture (Hartiala et al., 2017).

Some interest in searching risk factors for CAD has been targeted on the male mediated Y chromosome, since CAD shows a higher prevalence in males compared to females. Indeed, previous studies have suggested the higher CAD morbidity in males could, at least partially, be explained by Y-chromosomal genetic variation (Eales et al., 2019). Because Y-chromosomal haplogroups display differing distributions between eastern and western Finland (Lappalainen et al., 2006; Neuvonen et al., 2015), they could therefore also explain some extent of the geographical risk for CAD observed in Finland. Interestingly, the geographical risk for CAD is opposite to the CAD-linked Y-chromosomal haplogroup I1 frequency in Finland: haplogroup I1 displays a high prevalence in western Finland, where contrastingly, the morbidity for CAD is the lowest in Finland. Overall, the role of Y chromosome in complex disease is much debated, and further studies of the suggested link between Y-chromosomal haplogroup I1 and CAD would be required to validate the association.

2 AIMS OF THE THESIS

The role of Y-chromosomal haplogroups in complex disease remains poorly understood since the Y chromosome is often excluded from genome-wide association studies. Nevertheless, previous studies suggest Y-chromosomal haplogroup I1 carries a higher risk to coronary artery disease (CAD) in the UK (Eales et al., 2019). Since the Finnish population harbors a notable frequency of Y-chromosomal haplogroup I1 (Lappalainen et al., 2006; Neuvonen et al., 2015), the Finnish population provides an opportunity to further study the suggested association between Y-chromosomal haplogroup I1 and CAD. Interestingly, the Y-chromosomal haplogroup I1 is more frequent in western Finland, while CAD is contrastingly more common in eastern Finland. Previous studies of Finnish Y-chromosomal haplogroups have concentrated mainly on assessing Finnish population history (Lappalainen et al., 2006; Neuvonen et al., 2015). In addition, because of the low number of samples (N < 600) and poor coverage of sampling regions included in these previous studies, the complete regional distributions of Y-chromosomal haplogroups in Finland has not yet been characterized.

This thesis has two aims:

1. Determining Y-chromosomal haplogroups in Finland and characterizing their geographical distributions using genotyping array data from the FinnGen project Data Freeze 6 dataset.

2. Evaluating the association between Finnish Y-chromosomal haplogroups and CAD by logistic regression.

3 MATERIALS AND METHODS

3.1 STUDY SUBJECTS AND GENOTYPE DATA

3.1.1 FinnGen study subjects and ethics statement

The data for this study was acquired from the FinnGen project Data Freeze 6, and samples that were genotyped on FinnGen custom genotyping arrays were selected for this study. The dataset comprised genotyping array data for 81 576 Finnish males. The FinnGen is a nation-wide project launched in 2017, collecting and combining genomic and health care data of Finns. The project aims to collect biological samples from 500 000 participants, covering approximately 10% of the Finnish population by the end of 2023. The FinnGen project combines Finnish universities, the National Institute for Health and Welfare, the Finnish Red Cross Blood Service, biobanks, hospitals and pharmaceutical companies. Ultimately, the FinnGen project aims in identification of new therapeutic targets, diagnostics and treatment of numerous diseases. Currently the FinnGen project contains over 300 000 samples available for research use (https://www.finngen.fi/en 14.4.2021).

The FinnGen study protocol number HUS/990/2017 has been approved by the Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS). The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, THL/2364/14.02/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020) and Statistics Finland (permit numbers: TK-53-1041-17 and TK-53-90-20). The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 6 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001.

3.1.2 Sample genotyping

Genotyping of the samples 81 576 was performed at the Thermo Fisher genotyping service facility in San Diego. The samples were genotyped on two FinnGen genotyping arrays. Genotyping of 48 017 samples was performed on FinnGen Thermo Fisher Axiom custom array v1. The FinnGen chip v1 comprises in total 837 899 markers, of which 15 are Y-chromosomal markers. The remaining 33 559 samples were genotyped on FinnGen Thermo Fisher Axiom custom array v2. The FinnGen chip v2 comprises in total 655 973 markers, of which 606 are Y-chromosomal markers (https://www.finngen.fi/en/researchers/genotyping, 14.4.2021). The genotype calling was performed at the Institute for Molecular Medicine Finland (FIMM) and the marker positions were mapped to Human Reference assembly GRCh38. The Y-chromosomal genotypes were quality filtered separately from autosomal and mitochondrial genotypes by the FinnGen data management team.

3.2 Y-CHROMOSOMAL HAPLOGROUP CALLING

The Y-chromosomal genotype data was available in variant call format (VCF). The VCF file was recoded in PLINK v1.9 (Purcell et al., 2007) to PLINK-format. The genotype positions were converted from Human Reference assembly GRCh38 to GRCh37 using liftOver (Hinrichs et al., 2006) for further data analysis. Y-chromosomal haplogroup calling was performed using yHaplo (Poznik, 2016), which uses phylogenetically informative single nucleotide polymorphisms (SNPs) from the International Society of Genetic Genealogy (ISOGG) database to assess Y-chromosomal haplogroup lineages for the samples.

3.3 FILTERING HAPLOGROUP CALLS

Analysis on the Y-chromosomal haplogroups was performed in R version 4.0.2 (R Core Team, 2020) using tidyverse (Wickham et al., 2019) and data.table (Dowle and Srinivasan, 2020) R-packages. Individuals genotyped on FinnGen chip v1 (N = 48 017) were excluded from the analysis, since the 15 Y-chromosomal markers on the genotyping chip were not sufficient for assessing Y-chromosomal haplogroups for the samples (missingness rate for haplogroup calls 95%). From the remaining 33 559 samples genotyped on FinnGen chip v2, individuals with no successful haplogroup call (N =

538) and individuals with Y-chromosomal haplogroup AT-0 (N = 36) were excluded from the dataset, leaving a dataset of 32 965 samples.

The dataset was further filtered by using the results from a pre-assessed PCA produced by the FinnGen data management team. The PCA was performed for the dataset by merging European samples from the 1000 Genomes Project genotype data with 271 341 FinnGen samples, in order to assess ethnic outliers in the FinnGen data. Additionally, kinship analysis was performed to assess genetic relatedness in the FinnGen samples, and samples of degree 2 relatedness were listed as related. The samples that were considered Finnish and unrelated were listed as FinnGen inliers. This information of inliers was used in filtering the data, leaving 24 160 samples in the dataset.

3.4 ASSESSING REGIONAL HAPLOGROUP FREQUENCIES

Y-chromosomal haplogroup frequencies were calculated for the dataset using the output of yHaplo. The haplogroups given by yHaplo (e.g., R1bxxxx) were combined to represent haplogroup clades (e.g., R1b). The haplogroup clades representing a frequency of 1% or more in the dataset were classified as major haplogroups. The remaining haplogroups that were not classified into these major clades were grouped together into a group named “Others” for the further analysis.

The Y-chromosomal haplogroup dataset was combined with phenotypic information. The geographical origin of the samples was defined by birth region information. The birth regions included in total 21 areas, covering 19 Finnish regions (Uusimaa, Southwest Finland, Satakunta, Kanta-Häme, Pirkanmaa, Päijät-Häme, Kymenlaakso, Southern Karelia, Southern Savonia, Northern Karelia, Northern Savonia, Central Finland, Southern Ostrobothnia, Ostrobothnia, Central Ostrobothnia, Northern Ostrobothnia, Kainuu, Lapland, Åland), abroad and the area of ceded Karelia. The area of ceded Karelia consists of former eastern Finnish territories that were lost to the Soviet Union during the World War II. The representativeness of different Finnish regions in the dataset was assessed by comparing the number of individuals within the Finnish regions in the data with the population structure of Finland reported by Statistics Finland (https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html#Population%20data%20by%20region 10.1.2021).

The birth region information was used for calculating Y-chromosomal haplogroup frequencies for each of the birth regions individually. The regional haplogroup frequencies were visualized on a map

of Finland for 20 regions, excluding samples from Åland due to the low number of samples (N = 36). The analysis was conducted in R version 4.0.2 (R Core Team, 2020) using sp (Pebesma and Bivand, 2005) and mapplots (Gerritsen, 2018) R-packages. The geographical map of Finland was downloaded from GADM for level 2 and for Åland for level 0 (version 2.8, https://gadm.org/).

3.5 ASSESSING DIFFERENCES BETWEEN EASTERN AND WESTERN FINLAND

The samples were divided into northeastern (NE) and southwestern (SW) subpopulations based on their birth region location, to assess the differences in Y-chromosomal haplogroup frequencies between these areas. The geographical border between NE and SW areas was defined according to the reported population structure of the autosomal genome in Finland, which closely reflects the former political borderline of Treaty of Nöteborg from 1323 between Sweden and Novgorod (Kerminen et al., 2017). Lapland, Northern Ostrobothnia, Kainuu, Northern Savonia, Southern Savonia, Northern Karelia, Southern Karelia, ceded Karelia and Central Finland were categorized as the northeastern area (NE), whereas Central Ostrobothnia, Ostrobothnia, Southern Ostrobothnia, Satakunta, Pirkanmaa, Southwest Finland, Kanta-Häme, Päijät-Häme, Uusimaa, Kymenlaakso and Åland were categorized as the southwestern area (SW). Samples from abroad (N = 314) were excluded from the NE/SW classification. Y-chromosomal haplogroup frequencies were calculated for the subpopulations of NE and SW areas in R. The 95% confidence intervals (CI) for the haplogroup frequencies were calculated as

$$95\%\ CI = estimate \pm 1.96 * SE \quad (4)$$

The observed Y-chromosomal haplogroup frequencies within the whole dataset and within the NE and SW subpopulations were compared to similar earlier studies of Lappalainen et al. (2006) and Neuvonen et al. (2015). The quantity of each observed haplogroup was compared to the haplogroup quantities reported in earlier studies by 5-sample z-test in R version 4.0.2 (R Core Team, 2020).

3.6 INFERRING CHANGES IN REGIONAL HAPLOGROUP FREQUENCIES

The regional haplogroup frequencies were assessed by the birth years of the individuals to examine whether the Y-chromosomal haplogroup quantities had altered over time. The Y-chromosomal haplogroup frequencies for each major haplogroup were calculated annually by using birth years of

the individuals. The haplogroup frequencies were assessed on 12 regions, including Lapland, Kainuu, North Ostrobothnia, Ostrobothnia, North Savonia, North Karelia, Central Finland, South Savonia, South Karelia, Tavastia, Southwest Finland, and Uusimaa. Individuals born outside the range of 1925 to 1995 (N = 2148) were excluded from the analysis to improve the accuracy of the regional change estimation. In order to visualize the annual haplogroup quantities, the averaged haplogroup frequency estimates were predicted in R using local polynomial regression fitting (loess) and setting parameter α to 0.5. The haplogroup frequency estimates and their 95% confidence interval curves were visualized for each haplogroup in R version 4.0.2 (R Core Team, 2020).

3.7 CONSIDERING POPULATION STRUCTURE IN THE SAMPLES

Since population structure can be assessed by PCs derived from genotype data, the dataset was combined with eigenvalues of PCs 1 – 20, acquired from PCA performed by the FinnGen data management team. To assess whether the autosomal genome varied between the Y-chromosomal haplogroups, the distribution of PC values was assessed separately for individuals of each major haplogroup. The PC value distributions of each PC were compared by Kruskal-Wallis test to determine whether there were statistically significant differences on the PC value distributions between two or more haplogroups. The data was visualized by the first two PCs for each haplogroup. The analysis was performed in R version 4.0.2 (R Core Team, 2020) using ggplot2 (Wickham, 2016) and ggpubr (Kassambara, 2020) R-packages.

3.8 CLASSIFICATION OF CORONARY ARTERY DISEASE CASES AND CONTROLS

Clinical endpoints from the FinnGen project were used for the defining the CAD status for the samples (DF6, https://www.finngen.fi/en/researchers/clinical-endpoints). Individuals in the dataset were classified into CAD-cases and CAD-free controls using definitions from the study of Eales et al. (2019). CAD-free cases were classified as individuals with medical record of either myocardial infraction (I9_MI) or coronary revascularization procedures (I9_REVASC), and individuals whose death had been attributed to cardiac causes (I9_K_CARDIAC) in the death registry data. CAD-free controls were classified as individuals with no history of angina (I9_ANGINA), myocardial infraction, coronary revascularization procedures, and whose dead had not been attributed cardiac causes. Additionally, individuals with drug purchase history of CAD-medications were excluded from the

CAD-free controls using Anatomical Therapeutic Chemical (ATC) Classifications. Acetylsalicylic acid (B01AC06), glyceryl trinitrate (C01DA02), isosorbide mononitrate (C01DA14), isosorbide dinitrate (C01DA08) and nicorandil (C01DX16) were classified as CAD-medications.

Principally the CAD definition described above was used in the analyses. Although coronary atherosclerosis is the major cause of CAD, in rare cases CAD may be caused also by nonatherosclerotic pathways (Bastante et al., 2014). To evaluate if the constructed CAD endpoint reflected the most common cause of CAD, coronary atherosclerosis, FinnGen clinical endpoint for coronary atherosclerosis (I9_CORATHER) was used to divide individuals into cases and controls. The I9_CORATHER endpoint includes individuals with health or death registry data of acute ischemic heart disease, chronic ischemic heart disease, mechanical complication of coronary artery bypass and valve grafts, and presence of aortocoronary bypass graft.

3.9 ASSESSING THE ROLE OF FINNISH Y-CHROMOSOMAL HAPLOGROUPS TO CAD

The Y-chromosomal haplogroups with more than 1% frequency in the dataset were selected for association testing with CAD in comparison to other haplogroups. The Y-chromosomal haplogroup status of each individual was set as binary (one vs. all) to assess the effect of having one haplogroup over another. Association testing was performed by logistic regression, which can be used in describing the relationship between a discrete outcome variable and one or more covariates (Lewis and Knight, 2012). The logistic regression model can be written to denote the log odds for CAD as

$$log(Y) = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p. \quad (5)$$

where Y is the probability of having CAD, and $X_1 \ldots X_p$ indicate covariates and $\beta_1 \ldots \beta_p$ their effect estimates, respectively. The odds ratio for each covariate can be derived from the model as e^{β}.

For the association analysis, the p-value threshold was adjusted by Bonferroni correction for four haplogroups p = 0.05/4 = 0.0125. The association analysis was performed for each haplogroup status by logistic regression in R version 4.0.2 (R Core Team, 2020). The accuracy of the estimated results were evaluated by 95% confidence intervals (95% CI), calculated as

$$95\%\ CI = e^{\beta \pm 1.96 * SE} \quad (6)$$

The OR point estimates and their 95% confidence intervals were visualized in R version 4.0.2 (R Core Team, 2020) using ggplot2 (Wickham, 2016) R-package.

The effect size of each haplogroup was compared between the models using a one-sample z-test. The z-value was calculated for the effect estimates from two models as

$$z = \frac{(\beta 1 - \beta 2)}{\sqrt{SE1^2 + SE2^2}} \quad (7)$$

where β_1 and β_2 are effect estimates for the same haplogroup status from two separate regression models, and SE_1 and SE_2 their standard errors, respectively. The significance of the z-value was assessed by 2-sided p-value in R version 4.0.2 (R Core Team, 2020).

Power analyses were performed in to evaluate the potential of this study to find statistically significant effects with similar effect sizes reported in the study of Eales et al. (2019). The statistical power of a test describes the probability that when there is a true effect, the test statistic will reach the given p-value threshold. In a case-control association study, squared Wald´s test statistics z^2 follows a chi-square distribution with one degree of freedom given the non-centrality parameter, defined as

$$NCP = 2f(1-f)n\phi(1-\phi)\beta^2 \quad (8)$$

where ϕ is the proportion of cases, n number of samples, β the effect estimate, and f the minor allele frequency. Power was estimated by deriving the chi-square distribution of the non-centrality parameter with one degree of freedom with the given p-value threshold of 0.0125 in R version 4.0.2 (R Core Team, 2020).

3.9.1 Association between CAD and Y-chromosomal haplogroups in FinnGen

The association analyses were performed for each haplogroup status separately in four different models to compare the effect of Y-chromosomal haplogroups on CAD under different covariate combinations (Table 1).

First, analysis was carried out by replicating the study setting of Eales et al. (2019) by including age, BMI, smoking status, hypertension and PCs 1 – 20 as covariates in model 1. However, data for several covariates included in the UK Biobank study by Eales et al. (2019), including physical activity level, household income, completion of further education, employment index, alcohol intake

frequency and family history of heart disease were not available for this study and could not be included as covariates.

In model 2, the effect of including polygenic risk scores for CAD as a covariate was assessed. Polygenic risk scores describe the inherited susceptibility for a complex disease, usually calculated as a weighted sum of GWAS identified trait-associated alleles. Polygenic risk scores for CAD were computed by the FinnGen data management team, and the scores were based on GWAS summary statistics from the CARDIoGRAMplusC4D study (Nikpay et al., 2015). Covariates included in the model 2 were age, BMI, smoking status, hypertension, PCs 1 – 20 and CAD PRSs.

In model 3, the effect of PCs was assessed by excluding them from the regression model. Haplogroup status, age, BMI, smoking status and hypertension were included as covariates in model 3.

Since the sample size in these analyses 1 to 3 was limited by the missing information for BMI and smoking status for the majority of individuals (N = 12 264), further analysis was performed without these variables to maximize the statistical power in the analysis. In model 4, age, hypertension, PCs 1 – 20 and CAD PRSs were included as covariates.

Table 1 Summary of regression models 1 – 4. Aim, covariates, and the number of samples are listed on the table. Association analyses were performed separately between CAD and each Y-chromosomal haplogroup status.

Model	Aim	Covariates	N (cases / controls)
1	Replicate the study setting of Eales et al. (2019)	age, BMI, smoking status, hypertension, PCs 1-20	2877 / 5985
2	Infer the impact of CAD PRSs	age, BMI, smoking status, hypertension, PCs 1-20, CAD PRSs	2877 /5985
3	Infer the impact of PCs	age, BMI, smoking status	2877 / 5985
4	Maximize sample size and statistical power	age, hypertension, PCs 1-20, CAD PRSs	4602 / 16,524

The effect estimates for each haplogroup status were compared between the models by one-sample z-test, and the odds ratio point estimates and their 95% confidence intervals were visualized in R.

3.9.2 Analysis of other potential factors affecting the results

To infer if some other factors besides the covariates affected the association results, the associations were further validated in four models (Table 2). The following association analyses were performed using the covariates defined in model 4.

First, the effect of haplogroup N1c1 was assessed, since this haplogroup is mostly enriched to northeastern Finland, and was not included in the previous study of Eales et al. (2019). The impact of haplogroup N1c1 was assessed by excluding individuals of haplogroup N1c1 from the dataset, and performing association analysis between CAD and each of the remaining haplogroups in model 5.

Next, since Y-chromosomal haplogroups display differences in their geographical prevalence within Finland (Lappalainen et al., 2006; Neuvonen et al., 2015); the interest was to assess whether the association results depended on the geographical location of the samples. The dataset was divided into northeastern samples in model 6, and into southwestern samples in model 7, and logistic regression was performed independently on both datasets.

Moreover, since the major cause for CAD is coronary atherosclerosis, association was performed between Y-chromosomal haplogroups and coronary atherosclerosis, to assess if the CAD endpoint reflected coronary atherosclerosis. Samples were classified into cases and controls based on the FinnGen clinical endpoint definition for coronary atherosclerosis (I9_CORATHER) in model 8.

Table 2 Summary of regression models 5 – 8. Aim, type of alteration, and number of samples are listed on the table. Association analyses were performed separately between CAD and each Y-chromosomal haplogroup status.

Model	Aim	What was altered	N (cases / controls)
5	To infer impact of Y-chromosomal haplogroup N1c1	Excluding N1c1 individuals	2738 / 9975
6	To infer the association results in northeastern Finland	Including samples only of NE origin	2075 / 5543
7	To infer the association results in southwestern Finland	Including samples only of SW origin	2501 / 10,714
8	To infer if the CAD endpoint reflects the major cause of CAD, coronary atherosclerosis	Defining CAD by FinnGen clinical endpoint I9_CORATHER	4381 / 18,522

The results from the regression models 5 – 8 were compared to the unaltered dataset (model 4) and to the results reported by Eales et al. (2019). The odds ratio point estimates and their 95% confidence intervals were visualized in R and the effect sizes were compared between the models using a one-sample z-test.

3.9.3 Altering the reference haplogroup

Since the effect estimates were impacted by which haplogroups were included in the logistic regression analysis as reference, further analysis was conducted by altering the haplogroup status variable. The binary haplogroup status was altered to describe the comparison between two haplogroups (e.g. N1c1 vs. R1a) instead of comparing one haplogroup against all others. The following association analyses were performed using the covariates defined in model 4.

Logistic regression was performed between CAD and haplogroup N1c1 by comparing it separately to the remaining haplogroups (I1, R1a, R1b, Other). Likewise, logistic regression was performed for haplogroup I1 by comparing it similarly to the remaining haplogroups (N1c1, R1a, R1b, Other). The odds ratio point estimates and their 95% confidence intervals were visualized in R and the effect sizes were compared between the models by one-sample z-test.

4 RESULTS

4.1 HAPLOGROUP CALLING FROM GENOTYPING ARRAY DATA

From the total number of 81 576 samples, haplogroup calling was successful for 98% of samples genotyped with FinnGen chip v2, whereas only 5% of FinnGen chip v1 genotyped samples resulted in a successful haplogroup call (Table 3). Based on previous studies on Finnish Y-chromosomal haplogroups (Lappalainen et al., 2006; Neuvonen et al., 2015), the expectation was to observe haplogroups N1c1 and I1 in the dataset. However, these haplogroups were not observed in samples genotyped with FinnGen chip v1 due to the poor coverage of Y-chromosomal markers, thus the samples genotyped on FinnGen chip v1 were excluded from further analysis.

Table 3. Comparison on genotyping arrays. FinnGen chip v2 contains 606 Y-chromosomal markers, whereas FinnGen chip v1 contains only 15 Y-chromosomal markers, not covering major haplogroups in Finland. Samples from FinnGen chip v1 were excluded from this study.

Haplogroup	FG chip v2	FG chip v1
N1c1	60.4%	0%
I1	26.2%	0%
R1a	6.0%	0%
R1b	4.9%	74.9%
Other	2.5%	25.1%
Successful call	33,001	2601
Missing call	558	45,416

4.2 Y-CHROMOSOMAL HAPLOGROUPS IN FINLAND

For the dataset of 33 001 samples genotyped with FinnGen chip v2 resulting in a successful haplogroup call, 64 haplogroups were observed with yHaplo. After filtering the dataset, 24 160 individuals were captured as inliers, i.e., genetically unrelated Finnish individuals and 59 haplogroups remained in the dataset. These 59 haplogroups given by yHaplo were grouped to present major Y-chromosomal haplogroup clades. The most frequent haplogroup clades observed in the dataset (>0.5%) are listed in Table 4. Haplogroups N1c1, I1, R1a and R1b displayed frequencies of 5% or more, whereas the remaining haplogroups had a low frequency in the dataset (≤1%) and

were grouped together as one group named “Others” for further analysis. These five haplogroups and their frequencies are presented in Figure 5. Y-chromosomal haplogroups N1c1 and I1 were the most common haplogroups in the dataset, respectively found in 60.2% and 26.4% of the samples. The haplogroups R1a and R1b also were present at notable frequencies in the dataset, being found in 6% and 5% of the samples, respectively. As expected, all other haplogroups were present at considerably lower frequencies, showing a combined frequency of 2.4% in the dataset.

Table 4. The most frequent haplogroup clades observed in the dataset of 24 160 unrelated Finns. Haplogroup, frequency, haplogroup defining marker, and global enrichment of the haplogroup are listed on the table. Haplogroups N1c1, I1, R1a and R1b comprise 97.6% of the dataset.

Haplogroup	Frequency	Marker / rsid	Global enrichment
N1c1	60.18%	N-M46 / rs34442126	eastern Finland, northern Eurasia, Finno-Ugric speaking populations (Lappalainen et al., 2006)
I1	26.44%	I-M253 / rs9341296	western Finland, Scandinavia, central Sweden (Lappalainen et al., 2006; Neuvonen et al., 2015)
R1a	5.99%	R-M449 / rs17306692	eastern Europe (Underhill et al., 2015)
R1b	5.01%	R-M343 / rs9786184	western Europe, central Europe (Underhill et al., 2015)
I2	0.97%	I-M438 / rs17307294	Europe, Balkans (Rootsi et al., 2004)
E1b1	0.51%	E-P179 / rs16980621	southern Europe, northeastern Africa (Cruciani et al., 2004)

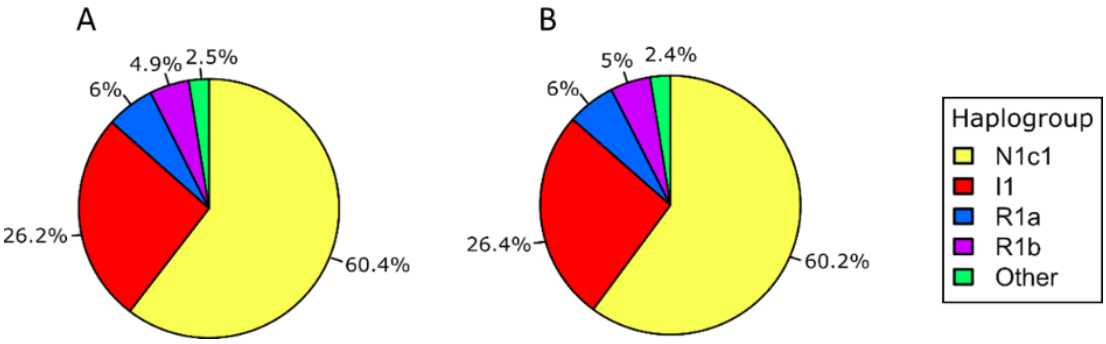

Figure 5. A) Major Y-chromosomal haplogroups observed in the full dataset of 33 001 males, and B) and in the filtered dataset of 24 160 unrelated Finnish males. Excluding related non-ethnic Finns from the dataset did not significantly alter the observed haplogroup frequencies for the most common haplogroups.

4.3 GEOGRAPHICAL ENRICHMENT OF Y-CHROMOSOMAL HAPLOGROUPS WITHIN FINLAND

After establishing frequencies for the Y-chromosomal haplogroups, birthplaces of the samples were used in determining how representative the data was of the overall Finnish population. From the 24 160 Finnish unrelated samples, information of the birth region was available for 24 119 individuals. Sample frequencies were calculated within each of the 18 Finnish regions (N = 23 074), excluding samples from abroad and ceded Karelia. The sample frequencies of each birth region were compared with the population structure of Finland from 2019 (https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html#Population%20data%20by%20region 10.1.2021) to infer the representativeness of different regions in Finland in the data. The correlation between the distribution of samples in the dataset and the regional distribution of the Finnish population (Pearson r = 0.94; 95% CI = 0.85-0.98; p = 2.5×10^{-9}) implied that the dataset comprehensively represented Finland.

The geographical distribution of the Y-chromosomal haplogroups was examined by calculating haplogroup frequencies within 20 birth regions, including 18 Finnish regions, abroad and ceded Karelia (N = 24 083) (Figure 6). Individuals from Åland were excluded due to the low number of samples (N = 36). The major Finnish haplogroups N1c1 and I1 accounted for most of the samples in Finland, but displayed altered patterns of enrichment depending on the geographical location as described in earlier literature (Lappalainen et al., 2006; Neuvonen et al., 2015). The frequency of Y-chromosomal haplogroup N1c1 was higher in northeastern regions compared to the southwest, whereas haplogroup I1 displayed higher frequencies in southwestern Finland compared to the northeast. For instance, in Kainuu (northeastern Finland), the observed frequencies for N1c1 and I1 were 78.0% and 12.4%, whereas in Satakunta (southwestern Finland), the frequencies were 48.6% and 37.5%, respectively.

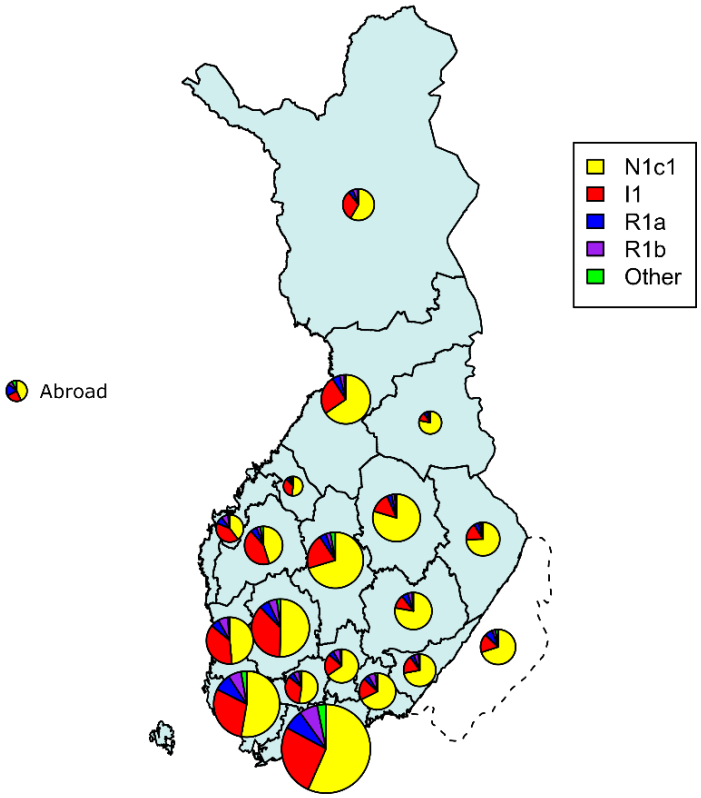

Figure 6. Y-chromosomal haplogroup frequencies by 20 birth regions, including 18 Finnish regions, abroad and ceded Karelia. The frequency pie charts in each region are scaled by the relative sample frequency of the largest pie (Uusimaa N = 4941).

4.4 DIFFERENCES BETWEEN EASTERN AND WESTERN FINLAND

Since differences are known to exist between individuals northeastern (NE) and southwestern (SW) Finland in the autosomal genome (Kerminen et al., 2017), health (https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index 1.4.2021), and Y-chromosomal haplogroups (Neuvonen et al., 2015; Lappalainen et al., 2006), the NE/SW differences were further assessed in the dataset. The geographical border between NE and SW areas was defined according to the reported population structure of the autosomal genome in Finland (Kerminen et al., 2017). Y-chromosomal haplogroup frequencies were calculated separately for northeastern (NE) and southwestern (SW) Finland (Figure 7). Samples from abroad were excluded from the NE/SW classification. Overall, 23 846 samples were included in the analysis, 15 012 samples representing the SW area, and 8843 samples the NE area, respectively.

NE Finland showed enrichment for haplogroup N1c1, present in 71.6% (95 %CI = 70.7%-72.5%) of the samples. Haplogroup N1c1 was also by far the most frequent haplogroup in SW Finland,

representing 53.8% (95% CI = 53.0%-54.6%) of the samples. The frequency of haplogroup I1 was 19.0% (95% CI = 18.2%-19.8%) in NE Finland, compared to 30.9% (95% CI = 30.2%-31.6%) in SW Finland. Haplogroups R1a, R1b, and Other displayed also a slightly higher prevalence in SW (15.3%) compared to NE Finland (9.5%).

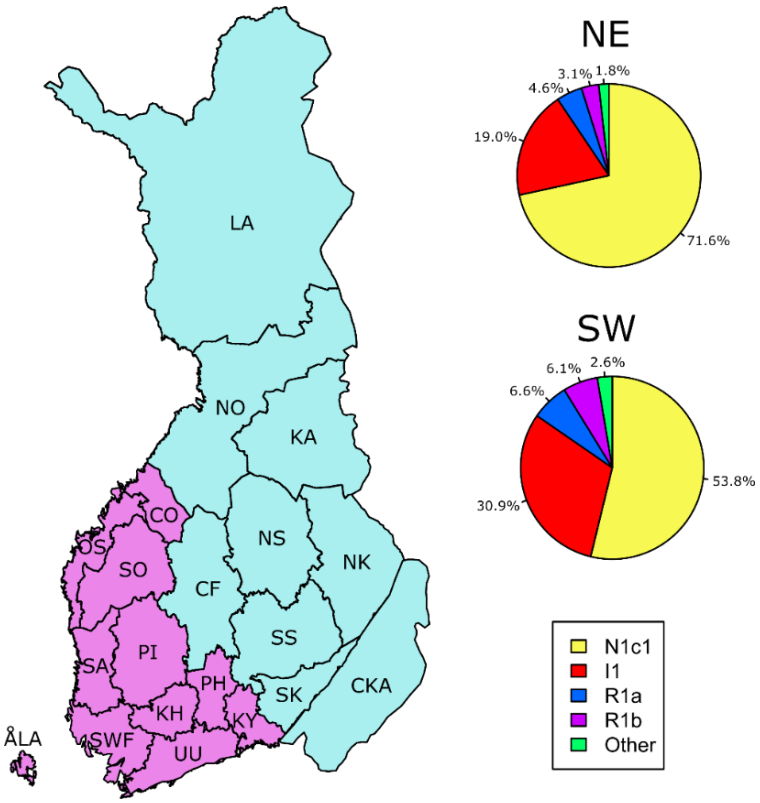

Figure 7. Y-chromosomal haplogroup frequencies illustrated for northeastern individuals (NE) and southwestern individuals (SW). The regions colored in cyan represent the NE area, whereas regions in purple represent the SW area.

The observed Y-chromosomal haplogroup frequencies in NE and SW Finland where compared to previously reported frequencies from Lappalainen et al. (2006) and Neuvonen et al. (2015) (Table 6). When considering NE Finland, the previously described Y-chromosomal haplogroup frequencies were similar to those found in this study ($p > 0.05$). However, the haplogroup frequencies observed in SW area contrasted with those reported earlier ($p < 0.05$). Previous studies reported enrichment of haplogroup I1 in SW Finland: respectively, 41.3% and 55.6% of SW samples represented haplogroup I1 in the studies of Lappalainen et al. (2006) and Neuvonen et al. (2015). However, the

observed frequency in SW Finland for haplogroup I1 was notably lower (31%) compared to the previous studies.

The inconsistencies in the SW frequencies were examined further by assessing the differences in sampling regions between this study and in the previous studies. Y-chromosomal haplogroup frequencies were calculated for SW area by including the same regions as incorporated by Lappalainen et al. (2006). Central Ostrobothnia, Ostrobothnia, Southern Ostrobothnia, Pirkanmaa, Satakunta and Southwest Finland were classified as the SW area, whereas Kanta-Häme, Päijät-Häme, Uusimaa, Kymenlaakso and Åland regions were excluded from the SW area. Repeating the analysis within the refined classification for SW area, the frequency of haplogroup I1 increased from 31% to 35.5% in the SW area. The observed Y-chromosomal haplogroup frequencies of SW Finland resembled those described by Lappalainen et al. (2006) ($p > 0.05$). The results suggest that the differences observed in the SW frequencies may result from the greater coverage of sampling regions included in this study compared to earlier studies.

Table 5. Frequencies of Y-chromosomal haplogroups N1c1 and I1 compared between this study and previous studies. For simplification, haplogroups R1a, R1b and Other are not presented in the table. The area of Finland includes samples combined from NE and SW regions. Area, haplogroup, number of samples, sample frequency, and 95% CIs are listed on the table.

Area	Haplogroup	Observed		Lappalainen et al. (2006)		Neuvonen et al. (2015)	
		N	% (95% CI)	N	% (95% CI)	N	% (95% CI)
Finland	N1c1	14397	60,4 (59,8-61,0)	312	58,2 (54,0-62,4)	289	49,5 (45,4-53,5)
Finland	I1	6312	26,5 (25,9-27,1)	155	28,9 (25,1-32,8)	242	41,4 (37,4-45,4)
SW	N1c1	8074	53,8 (53,0-54,6)	95	41,3 (34,9-47,7)	115	37,6 (32,1-43,0)
SW	I1	4632	30,9 (30,2-31,6)	95	41,3 (34,9-47,7)	170	55,6 (50,0-61,1)
NE	N1c1	6323	71,6 (70,6-72,5)	217	70,9 (65,8-76,0)	162	66,7 (60,7-72,6)
NE	I1	1680	19,0 (18,2-19,8)	60	19,6 (15,2-24,1)	58	23,9 (18,5-29,2)

4.5 INFERRING CHANGES IN REGIONAL HAPLOGROUP FREQUENCIES

To further examine whether the Y-chromosomal haplogroup frequencies had changed within the individual regions over time, haplogroup frequencies were calculated by birth years of the individuals. The haplogroup frequencies by time are presented for 12 regions in Figure 8.

The Y-chromosomal frequency estimates by year were most accurate in regions with the highest number of samples, such as Uusimaa (UU), whereas the frequency estimates were more inaccurate in regions with lower sample coverage, such as Kainuu (KA). Haplogroup N1c1 was the most frequent haplogroup within almost all regions through time. In Ostrobothnia, however, frequencies of haplogroups N1c1 and I1 were nearly equal from 1920 to 1975. Haplogroups R1a, R1b and Other were minor haplogroups through the time within all regions and did not display changes through the time. In eastern regions, such as south Savonia (SS) and north Karelia (NK) haplogroup N1c1 was by far the most frequent through time, whereas haplogroup I1 displayed lower frequencies. In the southwestern regions, such as Tavastia (TAV), Southwest Finland (SWF) and Uusimaa (UU), the haplogroup frequencies of N1c1 and I1 were more equal. In all of the graphs, the estimates were more inaccurate in the beginning and the end of the period due to statistical fluctuation. Overall, the Y-chromosomal haplogroup frequencies did not show intense changes over the time. The changes in haplogroup frequencies were considered as random variation, suggesting no adjustment for temporal changes of the haplogroups is required in subsequent analyses.

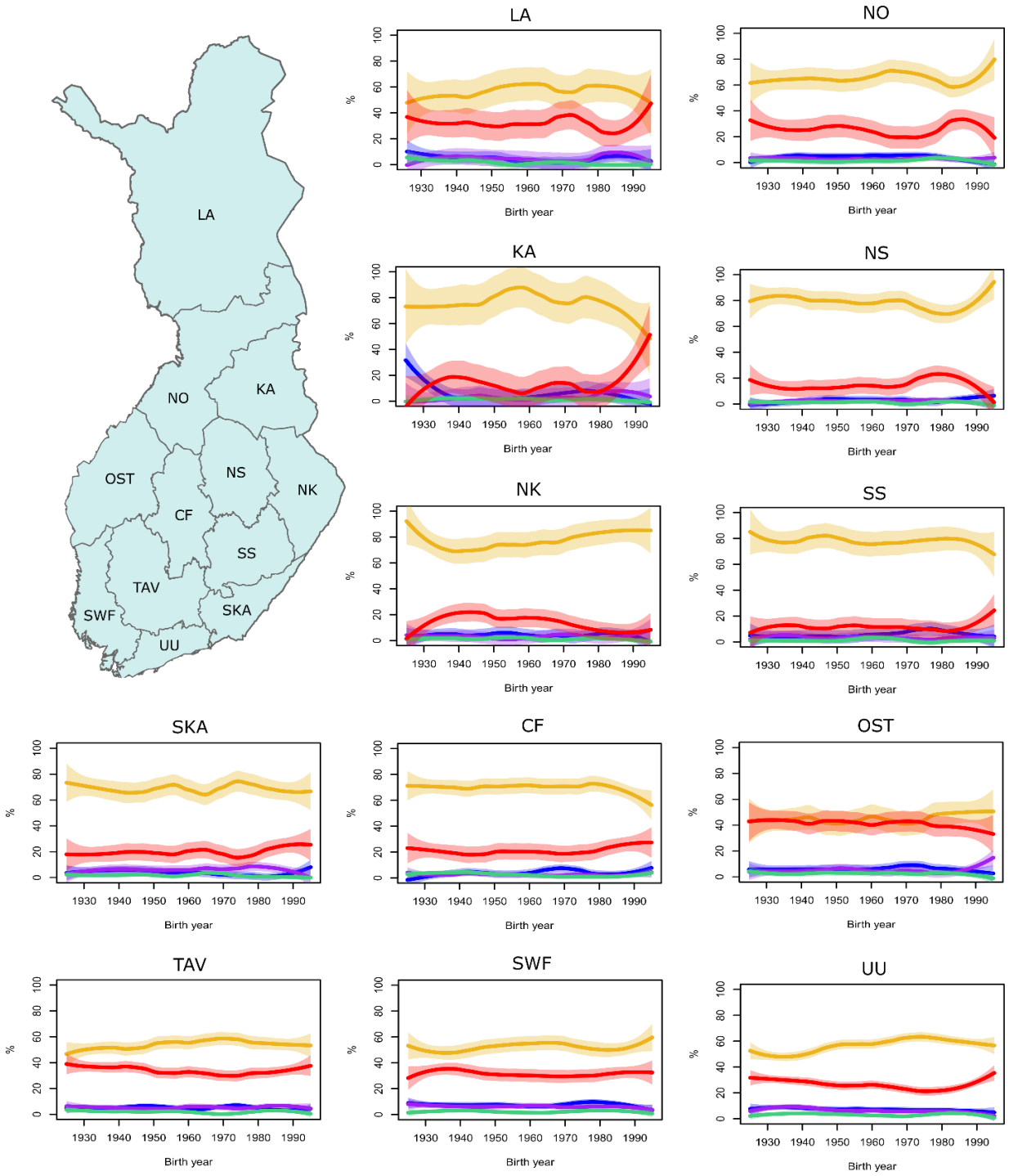

Figure 8. Y-chromosomal haplogroup frequencies from 1925 to 1995 for haplogroups N1c1 (yellow), I1 (red), R1a (blue), R1b (purple) and Other (green) by 12 regions.

4.6 POPULATION STRUCTURE IN THE SAMPLES

Since the autosomal genetic structure of Finns differs between NE and SW Finland (Kerminen et al., 2017), and the Y-chromosomal haplogroups display differing distributions between NE and SW Finland, correlation was tested between the autosomal genetic structure and Y-chromosomal haplogroups. The PC1 in particular is known to capture autosomal genetic differences between eastern and western Finns. Individuals were divided into five groups based on their Y-chromosomal haplogroup (N1c1, I1, R1a, R1b and Other). The distribution of PCs 1 to 20 were compared between these groups by Kruskal-Wallis test. The distribution of the values of first two autosomal PCs for each haplogroup is visualized in Figure 9.

A clear connection between PC1 and Y-chromosomal haplogroup N1c1 was observed in the analysis. Haplogroup N1c1 carriers had a greater quantity of negative PC1 values compared to other individuals, but also displayed a cluster of the positive PC1 values, similar to individuals of other haplogroups. The results thus indicate some correlation exists between the eastern Finnish autosomal genome and Y-chromosomal haplogroup N1c1.

When statistically comparing the PC value distributions between the haplogroups, significant differences were observed for PCs 1 to 18 (Kruskal-Wallis $p < 0.05$), implying correlation between Y-chromosomal haplogroups and the autosomal genome also on further PCs. The results suggested that adjustment for PCs 1-20 in subsequent association analyses is preferred to prevent confounding by population structure.

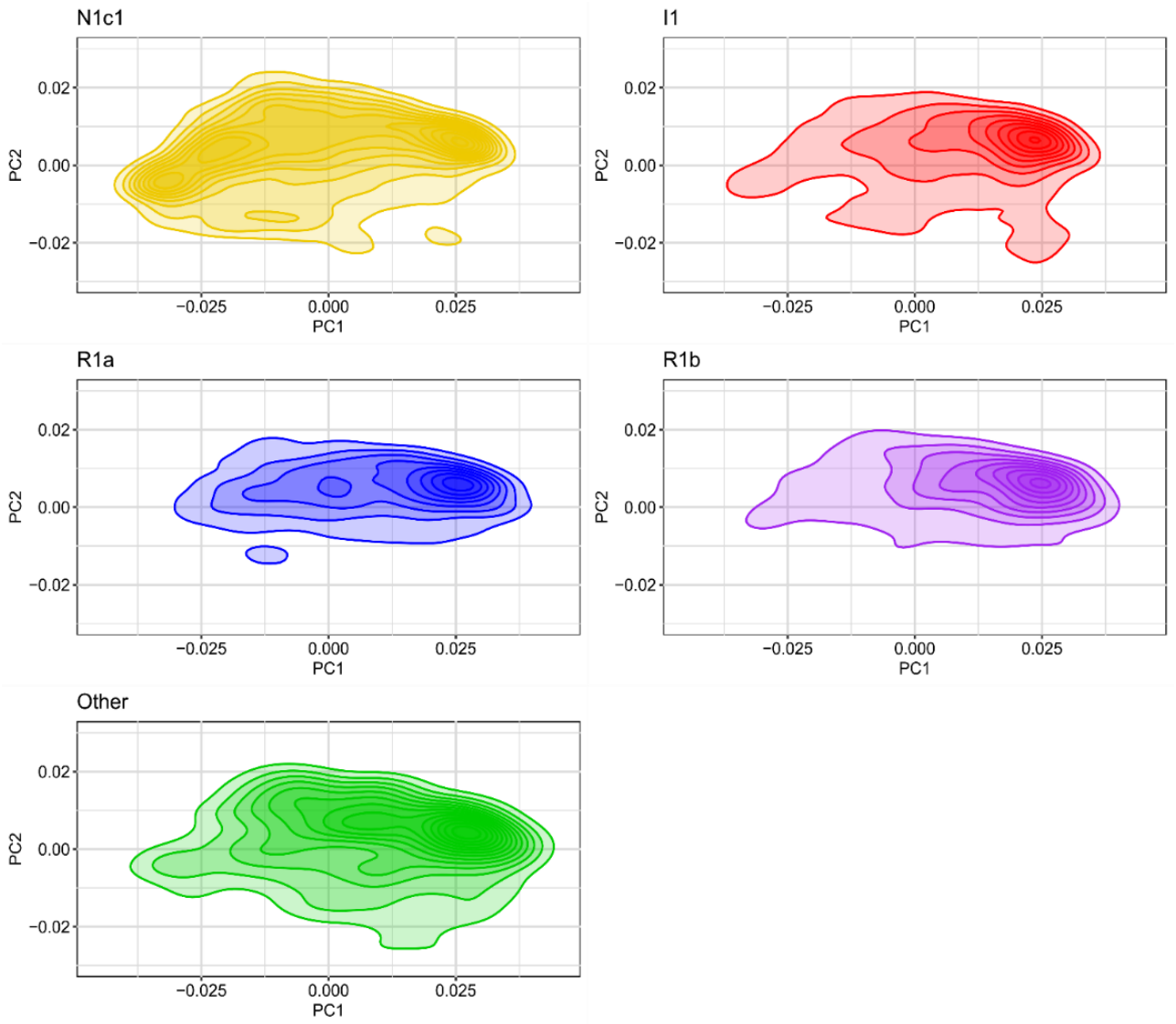

Figure 9. PC1 vs. PC2 visualized for carriers of different Y-chromosomal haplogroups. Carriers of Y-chromosomal haplogroup N1c1 differ on the first PC with a higher quantity of negative PC1 values compared to carriers of other haplogroups.

4.7 ASSESSING THE ROLE OF FINNISH Y-CHROMOSOMAL HAPLOGROUPS TO CAD

Since a previous study based on the UK Biobank data reported suggestive findings of haplogroup I1 increasing the risk for CAD in the British population (OR = 1.11, $p = 6.8 \times 10^{-4}$) (Eales et al., 2019), the association between CAD and Finnish Y-chromosomal haplogroups N1c1, I1, R1a and R1b was assessed. Logistic regression was performed between CAD and each binary haplogroup status in separate models (e.g. N1c1 vs. other haplogroups). To assess the robustness of the CAD associations, the role of covariates was evaluated by adjusting the association models by four different covariate combinations (see methods section 3.9.1).

First, the study setting of Eales et al. (2019) was replicated (model 1), next the effect of polygenic risk scores for CAD was assessed (model 2), and then the effect of PCs was assessed (model 3). Lastly, the sample size was maximized by excluding smoking status and BMI from the covariate set (model 4). The odds ratios and 95% CIs are presented for each haplogroup within these four regression models in Figure 10.

In the first model, haplogroups N1c1 and R1b displayed nominally significant associations with CAD. Haplogroup R1b displayed a risk increasing effect for CAD (OR = 1.26, p = 0.043), whereas the effect for haplogroup N1c1 was risk decreasing for CAD (OR = 0.87, p = 0.016).

The three remaining models were performed to assess if the results detected in the first analyses remained robust under different covariate configurations. Comparing the effect estimates for each haplogroup between models 1 – 4, no statistical differences were observed (z-test $p > 0.05$); implying most of the covariates explain independently a relatively small amount of the observed effect. The inclusion of polygenic risk scores for CAD in the second model appeared to have minor effects to the effect estimates. Similarly, the PC exclusion in the third model showed no significant impact on the effect estimates. The exclusion of smoking status and BMI in the fourth model did also not alter the effect estimates, but with a higher number of samples, the confidence intervals were decreased, thereby increasing the accuracy and statistical power of the analysis. With the increased sample size in the fourth model, a statistically significant association ($p < 0.0125$ corresponding for Bonferroni correction for four haplogroups) was observed for haplogroup N1c1 (OR = 0.91, p = 0.009). Overall, the results suggested for a protective effect for CAD in N1c1 haplogroup carriers.

Contrasting with the published results from the UK Biobank (Eales et al., 2019), haplogroup I1 did not display a significant association to CAD in any of the models (e.g. model 1: OR = 1.03, p = 0.7). However, the statistical power to detect a true association between haplogroup I1 and CAD with a similar effect reported in the UK (OR = 1.11) was suboptimal (64%) in model 1. Although statistical power was increased in model 4 (95% power for observing OR = 1.11 between I1 and CAD), no statistically significant association was observed between haplogroup I1 and CAD in the FinnGen dataset, suggesting haplogroup I1 does not associate with a higher risk to CAD in the Finnish population.

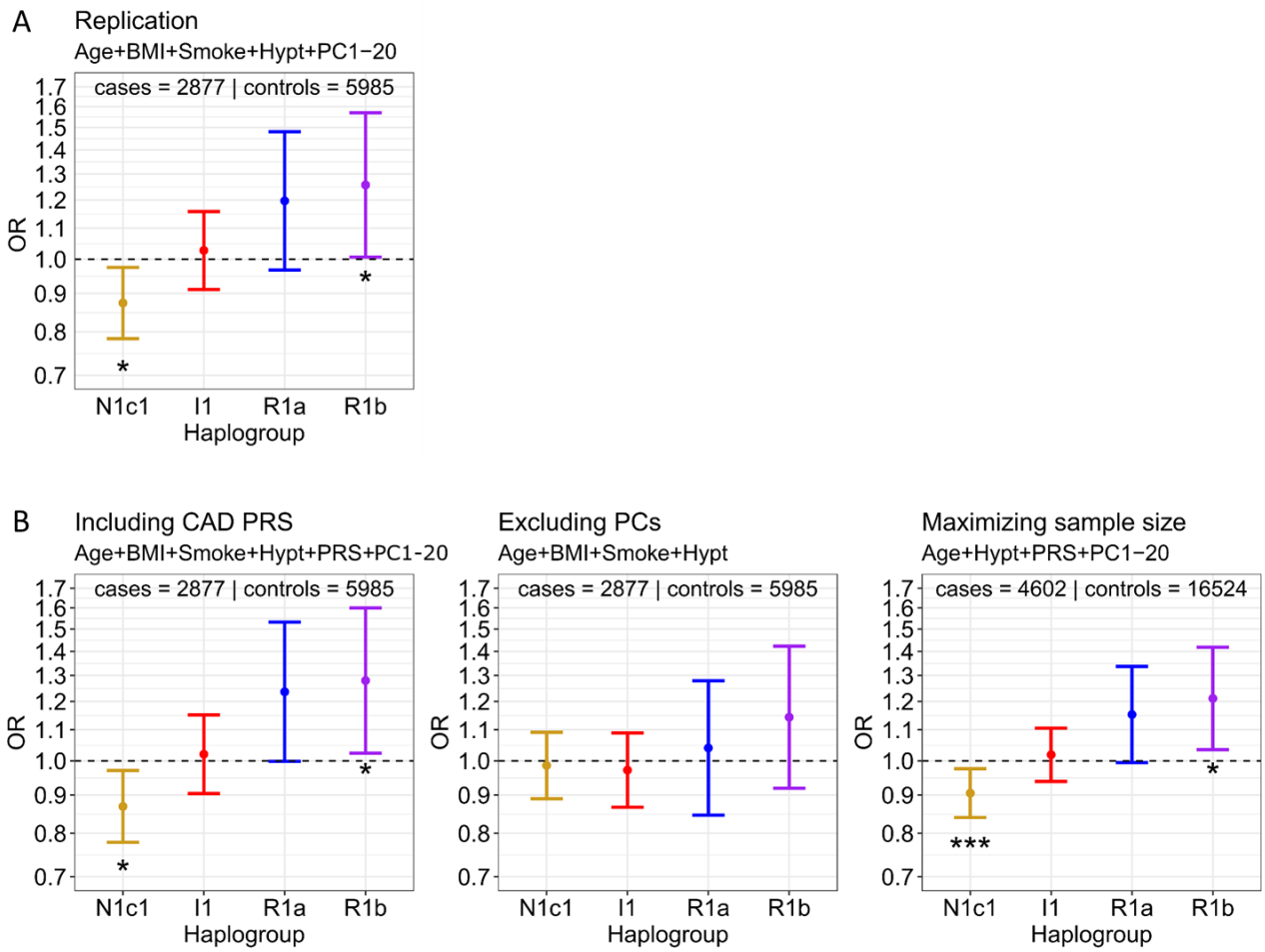

Figure 10. Analysis of the association of Y chromosome haplogroup to CAD using logistic regression models 1 – 4 including different covariates. A) Replication of the study setting described by Eales et al. (2019) (model 1) B) Altering the covariate combinations (models 2 – 4). Logistic regression was performed between CAD and each Y-chromosomal haplogroup status in separate models. Odds ratio, 95% CI, and significance are visualized on the graphs (* indicates $p < 0.05$ and *** indicates $p < 0.0125$).

4.8 FURTHER VALIDATION OF THE ASSOCIATION RESULTS

To test if some other factors besides the covariates had an impact on the haplogroup effect estimates, the associations were further inferred in four additional regression models (see methods section 3.9.2). Since covariates used in model 4 did not alter the effect estimates, but the larger sample size increased the accuracy of the results, the same set of covariates was used in the following association analyses.

First, the impact of haplogroup N1c1 was assessed on CAD by excluding it from the analyses, since haplogroup N1c1 had not been included the previous study by Eales et al. (2019) (model 5). Next, it was tested if the CAD associations could be replicated separately in northeastern samples (model

6) and southwestern samples (model 7). Additionally, to test whether the constructed CAD endpoint reflected the most common form of CAD, coronary atherosclerosis, the association between Y-chromosomal haplogroups and FinnGen clinical endpoint for coronary atherosclerosis (I9_CORATHER) was assessed. The odds ratios and 95% Cis for each haplogroup from the models 5 – 8 are presented together with the results from Eales et al. (2019) and the results covering the whole dataset (model 4) in Figure 11.

When considering models 4 – 8, the effect estimates did not change significantly for haplogroups N1c1, R1a and R1b between the models (one sample z-test $p > 0.05$). Conversely, haplogroup I1 displayed a significant difference in the effect estimate under model 5. When excluding haplogroup N1c1 carriers from the analysis, the neutral effect estimate for haplogroup I1 in model 4 (OR = 1.02, $p = 0.7$) changed to a risk decreasing effect for CAD (OR = 0.88, $p = 0.03$). Nevertheless, the exclusion of N1c1 carriers from the dataset did not alter the results for any other haplogroup besides haplogroup I1.

In models 6 and 7, the regional estimates in northeastern and southwestern samples did not introduce significant changes to the effect estimates for any of the haplogroups. However, the association between haplogroup R1b and CAD was most robust in northeastern Finland (OR = 1.42, $p = 0.018$).

Finally, in model 8, using the endpoint of I9_CORATHER instead of a broader endpoint of CAD, the effect estimates for each haplogroup remained similar, although with I9_CORATHER the disease associations did not reach statistical significance. The sample sizes were slightly different between the endpoints; CAD included approximately 5% more cases ($N_{cases} = 4602$) than I9_CORATHER ($N_{cases} = 4381$), that may partly explain the loss of significant association in the latter. Alternatively, the constructed endpoint for CAD might not fully reflect only coronary atherosclerosis.

The observed effect estimates were compared to the UK Biobank (UKB) results reported by Eales et al. (2019). The observed effect estimate for haplogroup I1 differed under two models from the UKB estimate: when excluding N1c1 carriers from the analysis, and when changing the endpoint definition to describe coronary atherosclerosis. However, most of the differences were observed for haplogroup R1b effect estimate, which differed in four models from the UKB estimate. Interestingly, the observed effects within all models for R1b were risk increasing (OR > 1), whereas in the UK the effect for R1b was reported as risk decreasing (OR < 1).

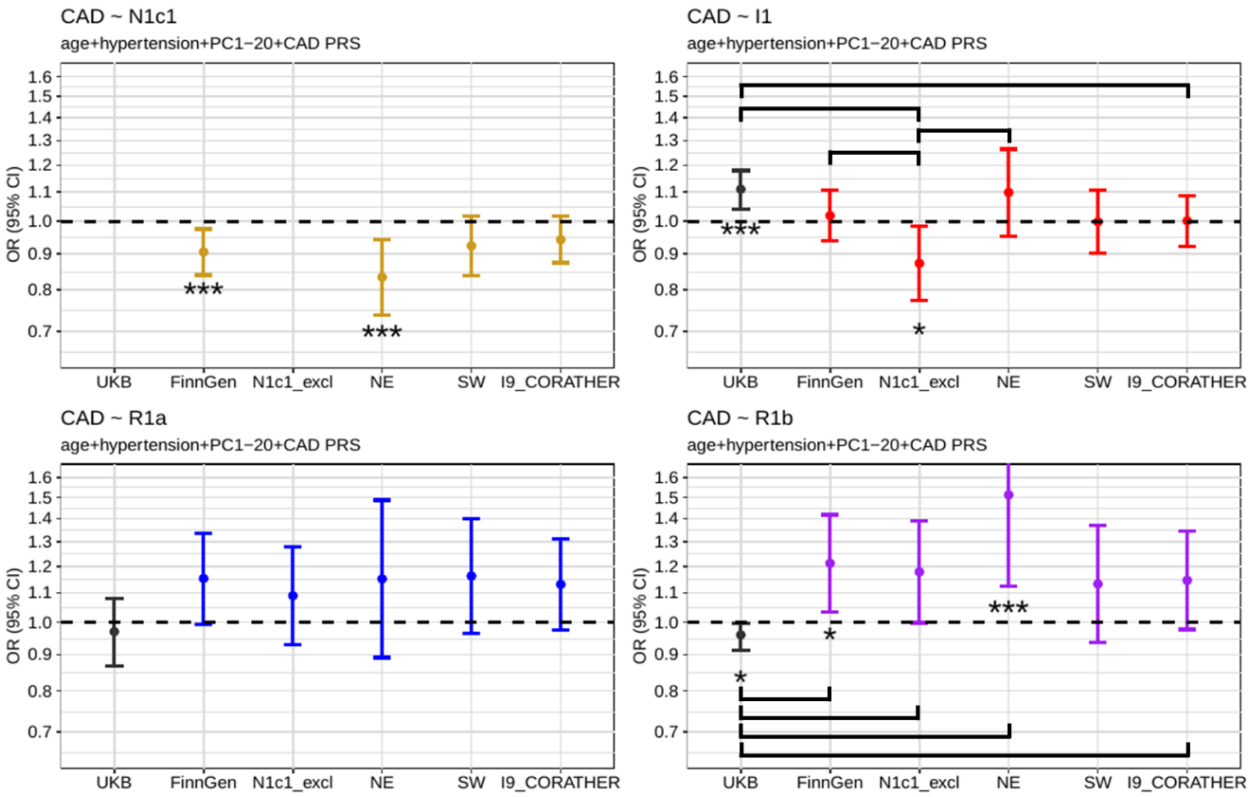

Figure 11. Associations between Y-chromosomal haplogroups and CAD compared between several models. Comparison of effect estimates between UK Biobank (UKB) (Eales et al. 2019), the FinnGen dataset (FinnGen) (model 4), without N1c1 carriers (N1c1_excl) (model 5), in northeastern Finland (NE) (model 6), in southwestern Finland (SW) (model 7) and by using I9_CORATHER instead of a broader CAD definition as an endpoint (model 8). Logistic regression was performed between CAD and each Y-chromosomal haplogroup status in separate models. Results are visualized each haplogroup separately: N1c1 (upper left), I1 (upper right), R1a (bottom left) and R1b (bottom right). For visual purposes, y-axis is truncated in the results for haplogroup R1b. Odds ratio, 95% CI, and significance are visualized on the graphs (* indicates $p < 0.05$ and *** indicates $p < 0.0125$). Black bars indicate significant ($p < 0.05$) differences between effect estimates from two models.

4.9 ALTERING THE REFERENCE HAPLOGROUP

Since the effect estimate for haplogroup I1 changed significantly, when excluding haplogroup N1c1 carriers from the analysis, further analyses were conducted by altering the haplogroup status variable to assess the impact of reference haplogroups on CAD associations. The binary haplogroup status was altered to describe the comparison between two haplogroups (e.g. N1c1 vs. R1a) instead of comparing one haplogroup against all others. Logistic regression was performed between CAD and haplogroup N1c1 by comparing it separately to the remaining haplogroups. Likewise, logistic regression was performed for haplogroup I1 by comparing it similarly to the remaining haplogroups. Logistic regression was performed between CAD and each pairwise haplogroup status in separate models (Figure 12).

The effect estimate for haplogroup N1c1 did not significantly change when the reference haplogroup was altered (z-test $p > 0.05$). Instead, the effect of haplogroup I1 displayed a significant change ($p < 0.05$), when comparing against haplogroup R1b. In addition, the effect between I1 and CAD was risk increasing when comparing against haplogroup N1c1, but risk decreasing when comparing to the remaining haplogroups R1a, R1b or Other.

Comparing the haplogroup I1 effect estimates to those reported by Eales et al. (2019), the results showed contrasting outcomes. The observed effect for haplogroup I1 was risk decreasing when comparing against haplogroup R1b (OR = 0.96), whereas in the same association was reported as risk increasing in the UK (OR = 1.11). Similarly, the effect for haplogroup I1 was risk decreasing when compared against R1a (OR = 0.89), whereas in the same association was reported to increase the risk for CAD in the UK (OR = 1.13). The contrasting results demonstrate that association results of Y-chromosomal haplogroups depend on the reference haplogroup set, and that the Y-chromosomal association results are not directly comparable between populations.

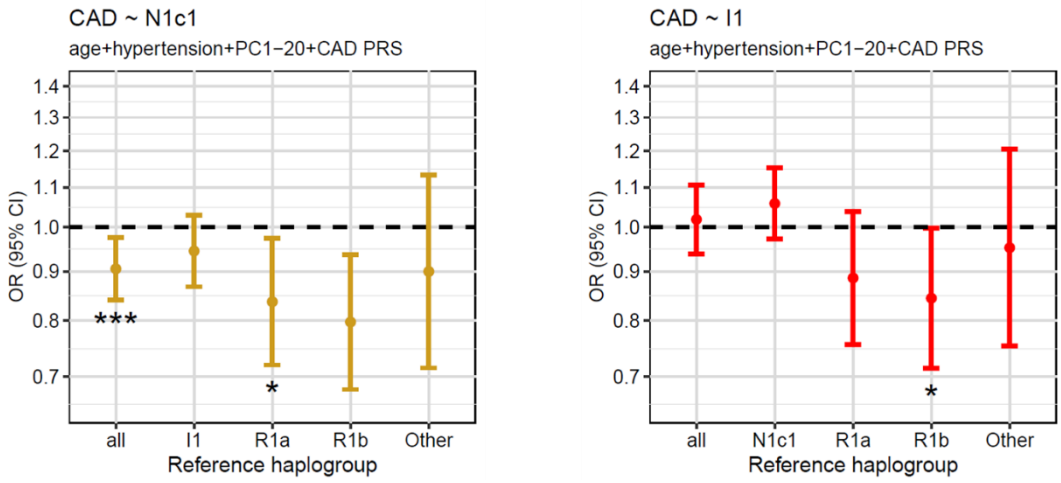

Figure 12. Associations between Y-chromosomal haplogroups and CAD compared between different reference haplogroups. Haplogroup N1c1 was compared to individual haplogroups in separate models (left). Similarly haplogroup I1 was compared to individual haplogroups (right). Odds ratio, 95% CI, and significance are visualized on the graphs (* indicates $p < 0.05$ and *** indicates $p < 0.0125$).

5 DISCUSSION

5.1 DERIVING Y-CHROMOSOMAL HAPLOGROUPS FROM GENOTYPING ARRAY DATA

This thesis characterized the prevalence of Y-chromosomal haplogroups in Finland using an extensive dataset of genotyping array data acquired from the FinnGen project, whereas previous studies have utilized considerably smaller datasets (N < 600) (Lappalainen et al., 2006; Neuvonen et al., 2015). Starting from 81 576 genotyped samples from the FinnGen Data Freeze 6 dataset, Y-chromosomal haplogroups were determined for 24 160 unrelated Finnish males. The samples were genotyped on two genotyping arrays, but haplogroups were successfully determined only for samples from one genotyping array. FinnGen chip v2 encompassed 606 Y-chromosomal markers, resulting in a successful haplogroup call for 98% of samples genotyped on that array. Contrastingly, FinnGen chip v1 encompassed 15 Y-chromosomal markers, resulting in a successful call only for 5% of the samples genotyped on the array. Therefore, samples genotyped on the FinnGen chip v1 were excluded from the dataset. Overall, the haplogroup calling highlighted that the haplogroups derived from each genotyping array should be assessed carefully to validate the calls and to make inferences about the prevalence of certain haplogroups in a population.

Genotyping the Y chromosome is often not comprehensive, since the Y chromosome contains a relatively small amount of common variation, and the genetic variations on the Y chromosome are mostly population specific. Moreover, the genetic marker coverage cannot be increased by imputation in a similar manner as in other genetic regions, since the Y chromosome lacks widely implemented imputation panels. Since the Y-chromosomal haplogroups are highly geographically clustered, each population having its specific combination of Y-chromosomal haplogroups (Jobling and Tyler-Smith, 2003), comprehensive Y chromosome genotyping requires a sufficient coverage of population specific Y-chromosomal markers on the genotyping array. For instance, haplogroup N1c1 is geographically enriched to Finland, whereas it is relatively rare in elsewhere (Lappalainen et al., 2006; Neuvonen et al., 2015). Thus, in most populations, the genotyping array is not required to cover the N1c1-defining marker to successfully determine Y-chromosomal haplogroups. Instead, when considering Finnish samples where 60% of men carry Y-chromosomal haplogroup N1c1, it is crucial to include the haplogroup N1c1-defining marker on the genotyping array. Such locally enriched haplogroups are not unique to Finland, and hence when using common genotyping arrays

for Y-chromosomal genotypes, it is essential to check the coverage of the arrays for the markers of interest before carrying out the studies.

This thesis demonstrated that selected genotyping array data can be used to successfully determine main Y-chromosomal haplogroups for a large number of participants. Nevertheless, the resolution of the Y-chromosomal haplogroups and subhaplogroups that can be inferred depends heavily on the genotyping array. The FinnGen chip v2 data used in this study, covered markers that provide a high resolution for subhaplogroups of haplogroup R1 (such as R1b1a2a1a2b), whereas no further subhaplogroups for haplogroup N1c1 could be detected in the dataset. The FinnGen genotyping chip is designed principally for genotyping Finnish individuals, but since the Y chromosome has not been considered in data collection, the Y-chromosomal markers provide insights for mostly cohorts with European ancestry, rather than providing a high coverage of markers seen frequently in the Finnish population. For future studies, it could be valuable to utilize genotype data derived from a genotyping array with markers for further Finnish Y-chromosomal subhaplogroups, especially for haplogroup N1c1, potentially providing more insight into the association results reported here. In addition, it would be valuable to obtain and use sequencing data for haplogrouping to validate the accuracy of the genotyping array data, and to derive a wider range of subhaplogroups in the Finnish population.

5.2 Y-CHROMOSOMAL HAPLOGROUPS IN FINLAND

In this thesis, the first aim was to determine Y-chromosomal haplogroups in Finland and their prevalence in different parts of the country by using genotyping array data from the FinnGen project. Y-chromosomal haplogroups were assessed successfully for 24 160 unrelated Finnish males, and four major haplogroup lineages (N1c1, I1, R1a and R1b) were observed in the dataset, representing 97.5% of the samples. Consistent with previous reports of Lappalainen et al. (2006) and Neuvonen et al. (2015), 86.6% of the samples fell into haplogroups N1c1 and I1. Globally, haplogroup N1c1 is most frequent in eastern Finland, northern Eurasia and Finno-Ugric speaking populations, whereas haplogroup I1 is enriched to western Finland, Scandinavia and especially to central Sweden (Lappalainen et al., 2006; Neuvonen et al., 2015). Of the samples, 11% fell into subhaplogroups of R1: 6% of the samples represented haplogroup R1a, and 5% haplogroup R1b,

respectively. Haplogroup R1a is geographically enriched to eastern Europe, whereas R1b is enriched to western and central Europe (Underhill et al., 2015).

Previous studies focusing on Finnish population history have described the uneven distribution of Y-chromosomal haplogroups in Finland (Kittles et al., 1998; Lappalainen et al., 2006; Neuvonen et al., 2015). In line with previous studies, the Y-chromosomal haplogroups displayed differences in their frequencies on regional level, especially between the southwestern and northeastern areas of Finland. Haplogroup N1c1 displayed a higher frequency in northeastern Finland, whereas haplogroup I1 was more frequent in the southwestern regions, consistent with earlier studies (Lappalainen et al., 2006; Neuvonen et al., 2015). However, contrasting the previous reports, the observed frequency for haplogroup N1c1 was higher (and respectively frequency of haplogroup I1 lower) in southwestern regions than previously described, implying the regional changes may not be as dramatic as previously reported. The observed discrepancies between this and earlier studies most likely result from the low number of samples (N < 600), and poor coverage of sampling regions included in the previous studies (Lappalainen et al., 2006; Neuvonen et al., 2015). By using the large-scale data from the FinnGen project, a more comprehensive estimate of the overall Y-chromosomal haplogroup frequencies within southwestern Finland was described in this project compared to previous studies.

The genetic division of Finns into northeastern and southwestern populations discussed above has been well described in earlier literature. The differences have been observed on both, autosomal genetic structure (Kerminen et al., 2017) and Y-chromosomal haplogroups (Lappalainen et al., 2006; Neuvonen et al., 2015). This study demonstrated that these components correlate with each other. Haplogroup N1c1, geographically enriched to eastern Finland, was related to the PC1 of autosomal genetic component that can be used to separate the Finnish population along the northeast southwest geographical axis. However, none of the other haplogroups displayed a similar enrichment pattern on PC1. Overall, correlation between Y-chromosomal haplogroups and the autosomal genome was observed on PCs 1- 18, highlighting the importance of considering population structure in genetic association studies involving the Y chromosome in Finland, and likely also in other populations where population substructure is present.

In addition to ancient population migrations leading to population bottlenecks in Finland, and potentially also to the existing East-West differences, other more recent large-scale migration events, affecting the genetic structure in Finland, are also well documented. During and after the

World War II roughly 400 000 individuals were relocated from eastern territories mostly to southern and western regions of Finland (Westerholm, 2002). In addition, on the second half of the 20th century, urbanization has shaped the population distribution within Finland (Kerminen et al., 2021). While these migrations are detectable in the autosomal genetics, these were not observed on Y-chromosomal haplogroup level, as the results imply that the Y-chromosomal haplogroup frequencies have remained relatively constant through the 20th century. One reason for not observing differences in Y-chromosomal haplogroup frequencies might be that Y-chromosomal haplogroups N1c1 and I1 are well presented throughout Finland, thus minor increase of either haplogroup is not observable in the dataset.

5.3 THE ASSOCIATION BETWEEN Y-CHROMOSOMAL HAPLOGROUPS AND CAD

This thesis demonstrated that Y-chromosomal haplogroups can be assessed from the FinnGen genotyping array data, allowing the classification of major haplogroups in Finland. The second aim was to evaluate genetic associations between Finnish Y-chromosomal haplogroups and CAD as an example of the possibility to assess the role of Y chromosome in complex disease. Logistic regression was performed between CAD and Finnish Y-chromosomal haplogroups N1c1, I1, R1a and R1b.

Several logistic regression models were performed to evaluate the robustness of the results. The results suggested the major Finnish Y-chromosomal haplogroup N1c1 has a risk-decreasing effect for CAD in the Finnish population when compared against all other haplogroups. In addition, the results implied for a nominal risk-increasing effect for haplogroup R1b in the analysis when similarly comparing against all other haplogroups.

The results did not significantly change when using varying covariate combinations, suggesting the covariates explain independently a relatively small amount of the observed results. However, since the information of smoking status and BMI were missing from the majority of individuals, the sample size and statistical power increased notably when these variables were excluded from the regression model.

When assessing the role of other potential factors affecting the results, such as geographical origin of the samples, the association results were mostly consistent between the models. One significant change was observed in haplogroup I1 effect size when excluding haplogroup N1c1 carriers from the dataset. Without haplogroup N1c1 in the dataset, the effect estimate for haplogroup I1 was risk

decreasing for CAD, whereas in other models the effect for I1 was risk increasing or neutral for CAD when compared against all other haplogroups. This observation highlights that the I1 association with CAD depends on the chosen reference group, thereby pointing to potential challenges in comparing haplogroup associations between populations with different haplogroup combinations.

When comparing the associations between the broad endpoint definition for CAD constructed based on Eales et al. (2019) and the FinnGen clinical endpoint for coronary atherosclerosis (I9_CORATHER), the effect estimates were highly similar. When using the endpoint definition by Eales et al. (2019), haplogroups N1c1 and R1b displayed nominally significant associations. However, when using the FinnGen clinical endpoint definition for CAD, including 5% fewer cases, none of the associations was statistically significant. The results could suggest the constructed endpoint for CAD might not fully reflect only coronary atherosclerosis.

The observed results were compared to the results reported by Eales et al. (2019). The study of Eales et al. (2019) reported an association between haplogroup I1 and an increased risk for CAD in the British population, whereas the same association for haplogroup I1 was not observed in the Finnish population in any of the models. However, the results demonstrated that the association results were strongly driven by the reference haplogroups used in the model. Eales et al. (2019) performed their association analysis using the haplogroups R1b, R1a, E, G, I1, I2 and J present in the British population, whereas in this study haplogroups N1c1, I1, R1a and R1b were used. Thus, the association models comparing one haplogroup against others could not directly be compared between these studies. Nevertheless, in further association models performed, the link between CAD and Y-chromosomal haplogroup I1 was risk decreasing when comparing only against haplogroup R1b (OR = 0.84, p = 0.047). Interestingly, the same association was reported as risk increasing in the UK (OR = 1.1, p = 0.0007). The contrasting results suggest the association results of Y-chromosomal haplogroups are not directly comparable between populations.

The contrasting results between this study and those reported by Eales et al. (2019) may result from multiple reasons. A potential concern in the study of Eales et al. (2019) is the adjustment of the association models only by PCs 1 to 5. In this study, it was demonstrated that Y-chromosomal haplogroups display correlation with the autosomal genome in Finland, thus the association analyses were adjusted by PCs 1 to 20. Given that the British population also harbors complex autosomal genetic substructure (Leslie et al., 2015), the association results by Eales et al. (2019) could in theory be affected by population structure, not captured by the first five PCs.

Moreover, the observed differences could also be driven by Y-chromosomal variants not captured in these studies. When considering Y-chromosomal haplogroups, they consist of several varying haplotypes that are not captured by the standard genotyping arrays. For example, differing variations may have accumulated to the Finnish and British I1 Y chromosome lineages during the 3500 years after haplogroup I1 founding point (Hallast et al., 2015). Thus, differing Y chromosome variations may exist in the same main Y-chromosomal haplogroup lineages between populations, leading to population specific effects.

Overall, comparing association results of Y-chromosomal haplogroups between populations remains challenging also since the correlation patterns with autosomal genetic profiles may vary. These results highlight that the association results of Y-chromosomal haplogroups are population specific, thus even when implying for a true association, the results might not replicate in another population.

The novel finding for haplogroup N1c1 decreasing the risk for CAD observed in this study remains intriguing, since the prevalence of haplogroup N1c1 is higher in northeastern Finland, where the incidence for CAD is higher compared to the southwest. The observed N1c1 association thus contrasts the epidemiological and autosomal findings implying higher CAD risk goes together with eastern Finnish ancestry. This association between N1c1 and CAD could not be compared to the previous study by Eales et al. (2019), since the haplogroup N1c1 is extremely rare in the British population.

5.4 POSSIBLE SOURCES OF ERROR

Several technical factors may have affected the results of this study. Firstly, technical issues may exist in sample genotyping. When considering the FinnGen chip v2 samples in Y-chromosomal haplogroup calling, 559 samples (2%) were not assigned into any haplogroup. The reason for these unsuccessful haplogroup calls might result from errors in genotyping, if certain markers were not observed for these samples due to technical errors. Furthermore, it could theoretically be possible, given the specific Y-chromosomal marker coverage requirement discussed earlier, that the 559 samples belong to a specific haplogroup that could not be identified. Thus, these 559 samples may require further attention, but overall, such small proportion of samples would not change the overview of the described haplogroups in Finland.

In the association analysis between Y-chromosomal haplogroups and CAD, the sample size was small (N = 8862) for a genetic association study when replicating the study setting from Eales et al. (2019). Thereby, the statistical power to find associations with effect sizes similar to those reported by Eales et al. (2019), was suboptimal (e.g. the statistical power to detect a true association for haplogroup I1 with OR 1.11 was 64%). In the model with the maximized sample size (N = 21 126) statistical power was increased (e.g. power to detect a true association for haplogroup I1 with OR 1.11 was increased to 92%). However, this model with maximized sample size was not adjusted for smoking status or BMI, since they were missing from a large number of participants. These are well-characterized risk factors for CAD, often used as covariates in studies of CAD. Reassuringly, however, the Y-chromosomal haplogroup effect estimates did not significantly change between the adjusted and unadjusted models, suggesting the impact of these covariates is uncorrelated with the haplogroup effect.

When considering the logistic regression models, the regional estimates for northeastern and southwestern subjects, as well as the estimates produced when excluding haplogroup N1c1 carriers, were conducted by taking subsets of the original dataset. In this approach, the effects of the covariates were estimated separately for each subset, potentially introducing more uncertainty in the estimates. A more refined way would have been to use and code these haplogroups and regions as factors in the regression model allowing the full dataset to be used.

When considering the use of CAD PRSs, previous studies have demonstrated that polygenic risk scores derived from the CARDIoGRAMplusC4D study (Nikpay et al., 2015) might display bias in the Finnish population (Kerminen et al., 2019). Although the results did not significantly change upon the inclusion of PRSs as covariates in the regression model, the use of these PRS should be considered carefully in future analyses.

Although the CAD classification described by Eales et al. (2019) was used in this study, the classification was not identical since some covariates used in the study by Eales et al. (2019) were not available in the dataset used here. For instance, the FinnGen dataset was lacking information for physical activity level, household income, completion of further education, employment index and alcohol intake frequency. In addition, the paternal or maternal histories of heart disease were not used in adjusting the model here. However, as was demonstrated, the results did not significantly change when excluding known risk factor variables for CAD (such as smoking status and

BMI), thus it seems unlikely that the difference in these covariates described here would have significantly affected the results.

Lastly, a source of error in large-scale biobank data might arise from ascertainment bias, since the samples might be collected from a higher count of disease-affected individuals than healthy ones. Nevertheless, in 2019, 18% of all deaths in Finnish males were attributed to CAD (https://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_002_en.html, 22.3.2021), and correspondingly the prevalence of CAD in the dataset was approximately 20%. Although these numbers are not directly comparable, the similarity suggests no significant ascertainment bias in the data.

5.5 FUTURE WORK

It would be of interest to further assess the reasons for the differences in Y-chromosomal haplogroup I1 association results in the British and Finnish populations. Differences could be assessed by comparing the Y-chromosome sequences, to establish whether different Y-chromosomal variants are present in the British and Finnish haplogroup I1 carriers. The differences could also be assessed by functional studies, for instance by comparing the Y chromosome gene expression profiles of British and Finnish I1 haplogroup lineages.

Another interesting follow-up would be to validate the finding for haplogroup N1c1 association with CAD. The association could be replicated by including data from a new release of the FinnGen project dataset (DF7, here Data Freeze 6 data was used). Additionally the association study could be replicated in another population where haplogroup N1c1 is present in sufficient quantities. The haplogroup N1c1 is reported to have a high frequency also in the Baltic populations, and for instance 33.9% in the Estonian males are carriers of haplogroup N1c1 (Laitinen et al., 2002), suggesting the association study could be replicated in the Estonian Biobank data (Leitsalu et al., 2015).

The access to whole sequence data would also enable a more detailed Y-chromosomal haplogroup classification of the Finnish Y chromosomes. Especially classification of haplogroup N1c1 into further subhaplogroups would be interesting, since no further subhaplogroups were detected with the genotyping array used for this study, and the potential disease impacts of these subhaplogroups thus remains unstudied. The whole sequence data would also be valuable in validating the accuracy of the haplogroups derived from genotyping array data.

Moreover, it would be interesting to further study other phenotypes besides CAD, to get a complete picture of the role of Y chromosome in human biology and its potential role in introducing sex differences in certain diseases. Sex and gender affect a wide range of biological functions, thus they have an impact on a wide range of diseases. In addition to cardiovascular diseases, many diseases are reported to exhibit sex differences, including for instance, neurological, pulmonary, and autoimmune diseases amongst others (Regitz-Zagrosek, 2012). The role of Y chromosome in disease could be studied by a phenome-wide association study. However, the current limitation is the number of samples, and larger datasets would be required for further studies to ensure sufficient statistical power.

6 CONCLUSIONS

In conclusion, this thesis provides a detailed characterization of Y-chromosomal haplogroups in Finland on regional level, and between northeastern and southwestern Finland. Using an extensive dataset 40 times larger compared to previous studies, this study demonstrates that the differences in the Y-chromosomal haplogroup prevalence between northeastern and southwestern Finland may not be as dramatic as previously suggested.

The genetic association study between Y-chromosomal haplogroups and CAD, first of its kind in Finland, demonstrates that a common Finnish Y-chromosomal haplogroup N1c1 associates with a lowered risk to CAD compared to carriers of other haplogroups. In addition, the results imply the Y-chromosomal haplogroup I1 is not associated with a higher risk to CAD compared to other haplogroups in the Finnish population, contrasting findings from British data. The results highlight the challenges in the interpretation of the results and demonstrate the need for further research in order to understand the debated phenotypic effects that the Y chromosome might have on complex disease.

Importantly, this study demonstrates that genetic associations of the Y chromosome can be studied utilizing large-scale biobank data, and that the Y chromosome might have a role in complex disease. However, the limitation in studying the genetics of Y chromosome arises from the lack of established Y-chromosomal marker coverage on genotyping arrays, thus available data on the Y chromosome is limited. Therefore, more effort should be placed in developing genotyping arrays with higher coverage of Y-chromosomal markers. Overall, I hope that this study motivates future studies to consider the inclusion of the Y chromosome more often in data collecting and genetic association studies.

7 ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors Taru Tukiainen and Jaakko Leinonen. Special thanks to both of you, Taru and Jaakko, for the professional help, genuine support, and helpful comments. Thank you Taru for offering the possibility to work on this exciting project especially during these challenging times. Thank you Jaakko for walking me through everything in this project, certainly none of this work could have been done without your help. I would also want to acknowledge my family and closest ones, thank you for being there for me at all times. I want to thank the FinnGen project community for providing the dataset, special thanks to Georg Brein, Awaisa Ghazal, Jarmo Harju, Elina Kilpeläinen, Priit Palta and Timo Sipilä. Moreover, thanks to all the anonymous participants of the FinnGen project. Thank you Sini Kerminen and Matti Pirinen for the interest towards this project, good discussions, and practical tips for the data analysis. Special thanks to Sini for sharing advice on plotting maps of Finland. Lastly, I would like to thank all other members in group Tukiainen, I truly wish to meet all of you in person after the coronavirus pandemic.

The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Celgene Corporation, Celgene International II Sàrl, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc., and Novartis AG. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi) and FinBB (https://finbb.fi/).

8 REFERENCES

1000 Genomes, P.C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526,* 68-74.

Armstrong, R.A. (2014). When to use the Bonferroni correction. Ophthalmic Physiol. Opt. *34,* 502-508.

Barros, B., Morais, M., Teixeira, A.L., and Medeiros, R. (2020). Loss of Chromosome Y and Its Potential Applications as Biomarker in Health and Forensic Sciences. Cytogenet Genome Res *160,* 225-237.

Bastante, T., Rivero, F., Cuesta, J., Benedicto, A., Restrepo, J., & Alfonso, F. (2014). Nonatherosclerotic causes of acute coronary syndrome: recognition and management. Current cardiology reports *16*, 543.

Baudat, F., Imai, Y., and de Massy, B. (2013). Meiotic recombination in mammals: localization and regulation. Nat. Rev. Genet. *14,* 794-806.

Bellott, D.W., Hughes, J.F., Skaletsky, H., Brown, L.G., Pyntikova, T., Cho, T., Koutseva, N., Zaghlul, S., Graves, T., Rock, S*., et al.* (2014). Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature *508,* 494-499.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E*., et al.* (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47,* D1005-D1012.

Colaco, S., and Modi, D. (2018). Genetics of the human Y chromosome and its association with male infertility. Reprod. Biol. Endocrinol. *16,* 14.

Cramp, L.J.E., Evershed, R.P., Lavento, M., Halinen, P., Mannermaa, K., Oinonen, M., Kettunen, J., Perola, M., Onkamo, P., and Heyd, V. (2014). Neolithic dairy farming at the extreme of agriculture in northern Europe. Proceedings. Biological Sciences *281,* 20140819.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E.B., Zaharova, B*., et al.* (2004). Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am. J. Hum. Genet. *74,* 1014-1022.

Dowle Matt and Srinivasan Arun (2020). data.table: Extension of `data.frame`. R package version 1.13.0. https://CRAN.R-project.org/package=data.table

Eales, J.M., Maan, A.A., Xu, X., Michoel, T., Hallast, P., Batini, C., Zadik, D., Prestes, P.R., Molina, E., Denniff, M.*, et al.* (2019). Human Y Chromosome Exerts Pleiotropic Effects on Susceptibility to Atherosclerosis. Arterioscler. Thromb. Vasc. Biol. *39,* 2386-2401.

Erzurumluoglu, A.M., Baird, D., Richardson, T.G., Timpson, N.J., and Rodriguez, S. (2018). Using Y-Chromosomal Haplogroups in Genetic Association Studies and Suggested Implications. Genes *9,* 45.

FinnGen clinical endpoints https://www.finngen.fi/en/researchers/clinical-endpoints referred 10.1.2021

FinnGen genotyping https://www.finngen.fi/en/researchers/genotyping referred 14.4.2021

FinnGen project https://www.finngen.fi/en referred 14.4.2021

Gerritsen Hans (2018). mapplots: Data Visualisation on Maps. R package version 1.5.1. https://CRAN.R-project.org/package=mapplots

GDAM version 2.8 https://gadm.org/ 10.1.2021

Godfrey, A.K., Naqvi, S., Chmátal, L., Chick, J.M., Mitchell, R.N., Gygi, S.P., Skaletsky, H., and Page, D.C. (2020). Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. Genome Res. *30,* 860-873.

Hallast, P., Batini, C., Zadik, D., Maisano Delser, P., Wetton, J.H., Arroyo-Pardo, E., Cavalleri, G.L., de Knijff, P., Destro Bisol, G., Dupuy, B.M.*, et al.* (2015). The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol. Biol. Evol. *32,* 661-673.

Hartiala, J., Schwartzman, W.S., Gabbay, J., Ghazalpour, A., Bennett, B.J., and Allayee, H. (2017). The Genetic Architecture of Coronary Artery Disease: Current Knowledge and Future Opportunities. Curr. Atheroscler. Rep. *19,* 6.

Helena Mangs, A., and Morris, B.J. (2007). The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Curr. Genomics *8,* 129-136.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F.*, et al.* (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34,* 590.

Hirschhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genetics in Medicine *4,* 45-61.

Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J.*, et al.* (2020). Ensembl 2021. Nucleic Acids Res. *49,* D884-D891.

Jobling, M.A., and Tyler-Smith, C. (2017). Human Y-chromosome variation in the genome-sequencing era. Nat. Rev. Genet. *18,* 485-497.

Jobling, M.A., and Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet. *4,* 598-612.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., and Hammer, M.F. (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. *18,* 830-838.

Kashimada, K., and Koopman, P. (2010). Sry: the master switch in mammalian sex determination. Development (Cambridge, England) *137,* 3921-3930.

Kassambara Alboukadel (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

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. (2021). Changes in the fine-scale genetic structure of Finland through the 20th century. PLOS Genetics *17,* e1009347.

Kerminen, S., Havulinna, A.S., Hellenthal, G., Martin, A.R., Sarin, A., Perola, M., Palotie, A., Salomaa, V., Daly, M.J., Ripatti, S., and Pirinen, M. (2017). Fine-Scale Genetic Structure in Finland. G3: Genes|Genomes|Genetics *7,* 3459.

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. (2019). Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. The American Journal of Human Genetics *104,* 1169-1181.

Kessler, T., Vilne, B., and Schunkert, H. (2016). The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. EMBO Mol. Med. *8,* 688-701.

Khramtsova, E.A., Davis, L.K., and Stranger, B.E. (2019). The role of sex in the genomics of human complex traits. Nat. Rev. Genet. *20,* 173-190.

Kim, S., and Misra, A. (2007). SNP Genotyping: Technologies and Biomedical Applications. Annu. Rev. Biomed. Eng. *9,* 289-320.

Kim, W., Yoo, T., Kim, S., Shin, D., Tyler-Smith, C., Jin, H., Kwak, K., Kim, E., and Bae, Y. (2007). Lack of association between Y-chromosomal haplogroups and prostate cancer in the Korean population. PloS One *2,* e172.

Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. *62,* 1171-1179.

Laitinen, V., Lahermo, P., Sistonen, P., and Savontaus, M.-. (2002). Y-Chromosomal Diversity Suggests that Baltic Males Share Common Finno-Ugric-Speaking Forefathers. Hum. Hered. *53,* 68-78.

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D.*, et al.* (2008). Correlation between genetic and geographic structure in Europe. Curr. Biol. *18,* 1241-1248.

Lappalainen, T., Koivumäki, S., Salmela, E., Huoponen, K., Sistonen, P., Savontaus, M., and Lahermo, P. (2006). Regional differences among the Finns: A Y-chromosomal perspective. Gene *376,* 207-215.

Leitsalu, L., Haller, T., Esko, T.õ, Tammesoo, M., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., Fischer, K., and Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol. *44,* 1137-1147.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J.*, et al.* (2015). The fine-scale genetic structure of the British population. Nature *519,* 309-314.

Lewis, C.M., and Knight, J. (2012). Introduction to genetic association studies. Cold Spring Harb Protoc. *2012,* 297-306.

Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T.*, et al.* (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. *10,* e1004494.

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schröder, R., and Stoneking, M. (2014). Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investigative Genetics *5,* 13.

Malakar, A.K., Choudhury, D., Halder, B., Paul, P., Uddin, A., and Chakraborty, S. (2019). A review on coronary artery disease, its risk factors, and therapeutics. J. Cell. Physiol. *234,* 16812-16823.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature Reviews Genetics *11,* 499-511.

Neuvonen, A.M., Putkonen, M., Översti, S., Sundell, T., Onkamo, P., Sajantila, A., and Palo, J.U. (2015). Vestiges of an Ancient Border in the Contemporary Genetic Diversity of North-Eastern Europe. PloS One *10,* e0130331.

Nikpay, M., Goel, A., Won, H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., Hopewell, J.C*., et al.* (2015). A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. *47,* 1121-1130.

Norio, R. (2003). Finnish Disease Heritage I: characteristics, causes, background. Hum. Genet. *112,* 441-456.

Ober, C., Loisel, D.A., and Gilad, Y. (2008). Sex-specific genetic architecture of human disease. Nat. Rev. Genet. *9,* 911-922.

Official Statistics of Finland (OSF): Causes of death [e-publication]. ISSN=1799-5078. 2019, Appendix table 1a. Deaths by underlying cause of death and by age in 2019, both sexes. Helsinki: Statistics Finland [referred: 22.3.2021]. Access method: http://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_001_en.html

Official Statistics of Finland (OSF): Causes of death [e-publication]. ISSN=1799-5078. 2019, Appendix table 1b. Deaths by underlying cause of death and by age in 2019, males. Helsinki: Statistics Finland [referred: 22.3.2021]. Access method: http://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_002_en.html

Official Statistics of Finland (OSF): Causes of death [e-publication]. ISSN=1799-5078. 2019, Appendix table 1c. Deaths by underlying cause of death and by age in 2019, females. Helsinki: Statistics Finland [referred: 22.3.2021]. Access method: http://www.stat.fi/til/ksyyt/2019/ksyyt_2019_2020-12-14_tau_003_en.html

Paracchini, S., Pearce, C.L., Kolonel, L.N., Altshuler, D., Henderson, B.E., and Tyler-Smith, C. (2003). A Y chromosomal influence on prostate cancer risk: the multi-ethnic cohort study. J. Med. Genet. *40,* 815-819.

Parker, K., Erzurumluoglu, A.M., and Rodriguez, S. (2020). The Y Chromosome: A Complex Locus for Genetic Analyses of Complex Human Traits. Genes (Basel) *11,* 1273.

Patel, R., Khalifa, A.O., Isali, I., and Shukla, S. (2018). Prostate cancer susceptibility and growth linked to Y chromosome genes. Frontiers in Bioscience (Elite Edition) *10,* 423-436.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

Poznik, G.D. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. bioRxiv 088716.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38,* 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.,I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81,* 559-575.

Regitz-Zagrosek, V. (2012). Sex and gender differences in health. Science & Society Series on Sex and Science. EMBO Rep. *13,* 596-603.

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

Rootsi, S., Magri, C., Kivisild, T., Benuzzi, G., Help, H., Bermisheva, M., Kutuev, I., Barać, L., Pericić, M., Balanovsky, O.*, et al.* (2004). Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. Am. J. Hum. Genet. *75,* 128-137.

Rootsi, S., Zhivotovsky, L.A., Baldovič, M., Kayser, M., Kutuev, I.A., Khusainova, R., Bermisheva, M.A., Gubina, M., Fedorova, S.A., Ilumäe, A.*, et al.* (2007). A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. European Journal of Human Genetics *15,* 204-211.

Salmela, E., Lappalainen, T., Fransson, I., Andersen, P.M., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M.L., Schreiber, S., Kere, J., and Lahermo, P. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS One *3,* e3519.

Sezgin, E., Lind, J.M., Shrestha, S., Hendrickson, S., Goedert, J.J., Donfield, S., Kirk, G.D., Phair, J.P., Troyer, J.L., O'Brien, S.,J., and Smith, M.W. (2009). Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. Hum. Genet. *125,* 281-294.

Sharp, A.J., Cheng, Z., and Eichler, E.E. (2006). Structural variation of the human genome. Annu. Rev. Genomics Hum. Genet. *7,* 407-442.

Shi, H., Dong, Y.L., Wen, B., Xiao, C.J., Underhill, P.A., Shen, P.D., Chakraborty, R., Jin, L., and Su, B. (2005). Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am. J. Hum. Genet. *77,* 408-419.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T.*, et al.* (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature *423,* 825-U2.

Statistics Finland, Population data by region https://www.tilastokeskus.fi/tup/suoluk/suoluk_vaesto_en.html#Population%20data%20by%20region referred 10.1.2021

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nature Reviews Genetics *20,* 467-484.

Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A., Rootsi, S., Honkola, T., Vesakoski, O., Atkinson, Q., Skoglund, P., Kushniarevich, A.*, et al.* (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biol. *19,* 139.

THL morbidity indexes in Finland https://thl.fi/en/web/thlfi-en/statistics/statistics-by-topic/morbidity/thl-s-morbidity-index referred 1.4.2021

THL morbidity indexes in Finland http://www.terveytemme.fi/sairastavuusindeksi/2016/maakunnat_html_profiili/atlas.html?select=01&indicator=i0 referred 1.4.2021

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazón Lahr, M., Foley, R.A., Oefner, P.J., and Cavalli-Sforza, L.L. (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. *65,* 43-62.

Underhill, P.A., Poznik, G.D., Rootsi, S., Järve, M., Lin, A.A., Wang, J., Passarelli, B., Kanbar, J., Myres, N.M., King, R.J.*, et al.* (2015). The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. Eur. J. Hum. Genet. *23,* 124-131.

Vartiainen, E. (2018). The North Karelia Project: Cardiovascular disease prevention in Finland. Global Cardiology Science & Practice *2018,* 13.

Voskarides, K., Hadjipanagi, D., Papazachariou, L., Griffin, M., and Panayiotou, A.G. (2014). Evidence for contribution of the y chromosome in atherosclerotic plaque occurrence in men. Genet. Test. Mol. Biomarkers *18,* 552-556.

Westerholm, John (2002). Populating Finland. Fennia - International Journal of Geography *180*, 123-140.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686.

Y Chromosome Consortium. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. *12,* 339-348.