Blood donor biobank suitability for identifying carriers with disease associated variants enriched in Finland

Eevaleena Vaittinen

Physiology and genetics
Master's thesis
Credits: 30

Supervisors:
Jonna Clancy
Satu Koskela
Irma Saloniemi

2.12.2022

Helsinki

Master's thesis

Subject: Physiology and genetics
Authors: Eevaleena Vaittinen
Title: Blood donor biobank suitability for identifying carriers with disease associated variants enriched in Finland
Supervisors: M.Sc. Jonna Clancy, Ph.D. Satu Koskela & Ph.D. Irma Saloniemi
Number of pages: 42 + 16 pages
Date: 2.12.2022

Biobanks collects samples from donors for medical research. Blood Service Biobank consists of samples from healthy enough blood donors, and it is assumed to be useful source for population and health studies because of the readily available sample collections. To understand better the actual usefulness of the blood donor biobank in genetic studies, in this study we have evaluated frequencies and geographical distributions of common and rare genetic variants in the Finnish population. The isolated population of Finland due to the bottleneck effect and the small founder population offers the opportunity to study genetic susceptibility variants that are rare in other populations. Finland has developed its own disease heritage, which consists of a group of recessive diseases that are more common in Finland compared to the rest of Europe. The population of a few thousand individuals settled mainly in the South and less in the East and the West. Another phenomenon that affected Finland's gene pool was the colonization of Northern/Eastern Finland from Southern Finland in the 15th-16th centuries by small families. The Finnish population has favorable characteristics for genetic research, such as general homogeneity, reduced diversity and increased linkage disequilibrium. The purpose of this Master of science project is to investigate how well blood donor Biobank can be used for identifying carriers with rare disease associated variants enriched in Finland

In this study 35 400 blood donors' genomic data was used for identifying the carriers of 51 disease susceptibility variants enriched in Finland among healthy blood donors. Allele frequencies for the variants were calculated based on number of heterozygotes and homozygotes. The allele frequencies of the SPR Blood Service Biobank were compared with two other Finnish datasets and one non-Finnish dataset. Blood donors' postal code information was combined with the genomic data and allocated to the right province based on the post code. This way we were able to compare variant distribution between 19 Finnish provinces

Results show that all of the investigated variants can be found among blood donors. Statistical tests showed significant differences in variant prevalence between East and West. Some of variants showed differences between provinces. Based on the study, it can be concluded that the Blood Service Biobank is well suited for finding rare disease associated variants among blood donors.

Key words: Biobank, blood donor, variant

Pro gradu -tutkielma

Pääaine: Fysiologia ja genetiikka
Tekijä: Eevaleena Vaittinen
Otsikko: Verenluovuttajabiopankin soveltuvuus Suomeen rikastuneiden tautialttius varianttien kantajien tunnistamiseen
Ohjaajat: FM Jonna Clancy, FT Satu Koskela & FT Irma Saloniemi
Sivumäärä: 42 + 16 sivua
Päivämäärä: 2.12.2022

Veripalvelun Biopankki koostuu verenluovuttajien näytteistä ja sen oletetaan olevan hyödyllinen lähde väestö- ja terveystutkimuksille helposti saatavilla olevien näytekokoelmien vuoksi. Ymmärtääksemme paremmin verenluovuttajabiopankin todellista hyödyllisyyttä geneettisissä tutkimuksissa, olemme tässä tutkimuksessa arvioineet yleisten ja harvinaisten geneettisten varianttien esiintymistiheyttä ja maantieteellistä jakaumaa Suomen väestössä. Suomen eristäytynyt populaatio pullonkaulaefektin sekä pienen perustajapopulaation vuoksi tarjoaa mahdollisuuden tutkia geneettisiä tautialttius variantteja, jotka ovat harvinaisia muissa populaatioissa. Suomeen on kehittynyt oma tautiperimä, joka muodostuu ryhmästä resessiivisiä tauteja, jotka ovat yleisempiä Suomessa verrattuna muuhun Eurooppaan. Muutaman tuhannen yksilön populaatio asettui pääosin etelään ja vähemmän itään sekä länteen. Toinen Suomen geenipooliin vaikuttanut ilmiö oli Pohjois-/Itä-Suomen kolonisaatio Etelä-Suomesta 1400–1500-luvulla pienten perheiden toimesta. Suomen väestöllä on geneettisen tutkimuksen kannalta edullisia piirteitä, kuten yleinen homogeenisyys, vähentynyt monimuotoisuus ja lisääntynyt kytkentäepätasapino. Tämän Pro gradu- tutkielman tarkoituksena on tutkia, kuinka hyvin verenluovuttajabiopankkia voidaan hyödyntää tunnistamaan Suomeen rikastuneiden harvinaisten tautialttius varianttien kantajia.

Tutkimuksessa käytettiin 35 400 verenluovuttajan genomitietoa. Genomitiedoista seulottiin 51 Suomeen rikastunutta harvinaista tautialttiusvarianttia tarkoituksena arvioida, kuinka hyvin variantteja löytyy terveestä verenluovuttaja populaatiosta. Varianttien hetero- ja homotsygoottien genotyyppien kantajat laskettiin, joiden avulla saatiin laskettua alleelifrekvenssit varianteille. SPR Veripalvelun Biopankin alleeli frekvenssejä verrattiin kahteen muuhun suomalaiseen populaatioon ja yhteen ei-suomalaiseen populaatioon. Verenluovuttajien postinumerotiedot yhdistettiin genomitietoihin ja jaettiin postinumeron perusteella oikeaan maakuntaan. Näin pystyimme vertailemaan varianttijakaumaa Suomen 19 maakunnan välillä.

Tulokset osoittavat, että kaikki tutkimuksessa käytettävät variantit löytyvät verenluovuttajista. Tilastolliset testit osoittivat merkittäviä eroja varianttien levinneisyydessä idän ja lännen välillä. Jotkut muunnelmat osoittivat korkeampia frekvenssejä maakuntien välillä. Tutkimuksen perusteella voidaan päätellä, että Veripalvelun Biopankki soveltuu hyvin harvinaisten tautialttius varianttien etsimiseen verenluovuttajista.

Avainsanat: Biopankki, verenluovuttaja, variantti

CONTENT

1. Introduction

1.1. Biobank

Biobanks collect biological samples (blood cells, DNA) and associated information (genomic and phenotype data) from donors for medical research and to promote health care. Biobanks are also a source for developing new medicine and treatment. Biobank samples and their data can be used for studying the etiology of diseases and validate new diagnostic methods. The biobank material can also be used for developing personalized medicine which aims to prevent, better diagnose, and treat patients individually. Widely collected samples can be used for research purposes efficiently, idea is to work as a functional unit to ease the access of researchers to high quality samples. Biobanks, which collect patients material, are focused on collecting specific diseased samples (Suomen Biopankit 2022).

There are 11 biobanks in Finland and the Finnish Red Cross (FRC) Blood Service Biobank is the only one that collects samples from blood donors. The Blood Service Biobank was established in summer 2017. Blood Service biobank, in which samples from blood donors are collected and saved, is assumed to be useful resource for population and health studies due to the existing and readily available sample collections. Blood donors are a highly selected group on the basis of health questionnaire, and the donor population is considered healthy as they are qualified to donate blood (Raivola et al 2019). Biobank samples are collected only in connection with blood donation. This means that healthy enough individuals are represented as the blood donor group and the material collected from blood donors can be used in research to represent a healthy control group This is how the Blood Service Biobank differs from other biobanks in Finland, as they collect samples from diseased individuals. The Blood Service Biobank provides samples and data from blood donors to help research aiming at preventing illnesses and identifying the pathological processes involved, the goal being to promote public health. Currently FRC Biobank has 35 400 blood donors' genotype data. Personal data from donors is kept secure under Biobank Act (688/2012) which purpose is to protect donor's rights. For processing personal data, Biobank follows the EU General Data Perfection Regulation (GDPR). The personal

data is collected from donors who have given biobanking consent in writing and it is optional. After signing the consent form, personal data is stored and can be used to enable scientific research (Blood Service 2021). The donor can review their personal data by filling in the Requested form or their personal data can be removed from Biobank by donor's request. If the donor sample has already been disclosed to a research project, the data will not be released for future research projects. To understand better the actual usefulness of the blood donor biobank in genetic studies, in this study we have evaluated frequencies and geographical distributions of common and rare genetic variants in the Finnish population.

1.2. Finnish population genetic background

The Finnish populations' unique genetic background offers ideal material for genetic studies. An isolated population such as the Finns provides a possibility to study genetic variants with effect on disease susceptibility and that are rare in other populations. As a result of the bottleneck in Finland approximately 120 generations ago and small founder population size, estimated 3 000-24 000 individuals, have affected on the Finnish disease heritage which consists of a set of recessive diseases that are common in Finland (Nevanlinna 1972). Around 4000 years ago migratory wave of eastern Uralic speakers have assumed to affect today's Finnish gene pool (Peltonem et al. 1995). Since the end of the last ice age, Finland has been constantly but sparsely populated indicated by archeological evidence (Palo et al. 2009) and the earliest signs of human activity post glacial era in the present-day Finland date back approximately 19 000 years. The earliest postglacial signals of human activity in the geographical area of present-day Finland date back approximately 10 900 years (Haggren et al. 2015). A few thousand individual population settled mostly in the south and lesser in the east and west. Another phenomenon affecting on the Finnish gene pool was the colonization of northern/eastern Finland (late settlement area, LAS) in the 15^{th}-16^{th} century from southern Finland (early settlement area) by small families (Kere 2001). The archeological materials along the Finnish coast during the Bronze age also shows notable Scandinavian influence and inland shows mostly Eastern features trough trading (Haggren et al. 2015).

The size of the Finnish population was only around 50 000 in the 12th century and reached 250 000 in the 16th century, inhabitation mainly concentrating in the coastal areas (Sajantila et al. 1996). In the 16th century the internal migration movement started in Finland from a smaller southeastern area to the middle, western and lastly northern and eastern parts of the country when the increasing population created pressure to culture more land. The Swedish Crown increased taxation which also had a major impact in the internal migration and development of administrative infrastructure reaching all the way to the back of Finland. Inhabitation of wilderness was preferred during the regime on the Swedish King Gustavus Vasa (1523-1560). During this regime the major recourse for later genetic studies was created which was church records. Records had information of births, deaths, marriages, and family movements providing reliable source of genealogical information since major of the population (>90%) belonged to the Evangelic Lutheran State Church. Almost entire Finland was inhabited sparsely but permanently by the end of the 17th century (Peltonen et al. 1999).

The reasons for the isolation of the Finnish population are caused by the geographical location of Finland, surrounded by the Baltic Sea from the south and west and the Arctic Ocean from the north. The geopolitical location between Sweden and Russian has also affected the isolation (Sajantila et al. 1996). Besides the isolation due the Nordic position, religious and language boundaries have also had impact on Finland's isolation and that way has affected enrichment of disease associated variants (Norio et al. 1973). The unique genetic structure of the Finnish population has had major effect on the research of rare and common diseases and their genetics as well as on health care practices. (Kääriäinen et al. 2017).

Compared to the early population genetic studies of Finns, the later results have shown that Finns are part of the genetic continuity between mainland Europe and Uralic -speaking populations from Siberia (Tambets et al. 2018). Genomewide data has shown genetic division between East and West (Salmela et al. 2008; Kerminen et al. 2017). A strong genetic borderline between Western and Eastern Finland have been shown by Y-chromosomal studies (Lappalainen et al. 2006). The Finnish population has advantageous features for genetic population studies in its genetic architecture such as overall

homogeneity, diminished diversity and increased linkage disequilibrium. Group of 20 rare inherited diseases, called the Finnish disease heritage, were initially described in 1973 as they are more frequent in Finland than in other populations in the world (Norio et al. 1973). Later the number of diseases increased in the group and today it includes more than 33 typically recessive diseases (De La Chapelle & Wright 1998). The founder effect means when a new population is formed from a small number of founding individuals and the populations' poor distinctive allele patterns causes decreased genetic diversity (Norino & Löytönen 2002). Recently reported marked reduction in Y chromosome diversity in Finnish population compared with other populations showed genetic evidence for Finnish population bottleneck (Sajatila et al. 1996). The enrichment of almost 40 rare diseases shows evidence of the founder effect as well as longer regions of linkage disequilibrium (LD) (Nevanlinna 1927). Diseases enriched in Finland are not evenly distributed. However, patients' birthplaces do not show any core areas of these disease genes due to the internal migration in Finland after the second world war. The birthplaces of grandparents show origins of the disease associated gene variants as the majority of the generation has been born in the countryside. (Norio & Löytönen 2002).

1.3. FinnGen-research project

FinnGen-research project is a public-private partnership project that collects samples and implements genotyping of about 500 000 Finnish individuals, which comprises 10% of the population. The aim is to study the genetic background of common diseases. The project combines imputed genotype data generated from the Finnish biobanks' samples and digital health record data from the health registries in Finland (FinnGen n.d.). Genotype imputation is a method to inference unobserved genotypes by using known haplotypes in a population (Scheet & Stephens 2006). Project aims to make Finland a pioneer in individual healthcare and biomedicine and to bring new treatments and heath innovations to the use of Finns. The FinnGen-project involves seven regional and three national biobanks, which are owned by hospitals, universities, and other research organizations (Finnish institute for health and welfare, FRC Blood Service) and the project is the first collaboration between Finnish biobanks (FinnGen 2022a). In accordance with the Biobank act and their own rules,

biobanks decide which samples and data will be provided for FinnGen-project. The genomic data is returned to biobanks after and will be available for all biobank research in the future. FRC Biobank blood donor genomic data is genotyped by FinnGen and returned back to Biobank. FinnGen uses reference allele frequencies; gnomAD-FIN and gnomAD-NFSEE (non-Finnish-Swedish-Estonian), because of the large-scale migration in the 20th century from Finland to Sweden and many chromosomes from Swedish sequencing studies shows recent Finnish origin. Estonia is likely to share same elements (Kurki M et al. 2022). Same reference allele frequencies are used in this study. FinnGen-project is funded by Business Finland and participating pharmaceutical companies.

1.4. Study variants

In this study the 51 rare disease associated variants enriched in Finland (Table 1.) will be investigated in the collection of blood donor samples. These variants were identified by Genome-wide association studies (GWAS) by FinnGen-project (Kurki M et al. 2022). Frequencies and geographical distributions of the 51 single nucleotide polymorphisms (SNP), such as chronic lower respiratory diseases, reported to be enriched in Finland and associated with disease susceptibility (Kurki et al. 2022), are calculated from the FRC Blood Service Biobank genomic material. The 51 variants were chosen as coding variants enriched over two-fold in Finns over Non-Finnish-Swedish-Estonian European populations in gnomAD that are significantly associated in GWAS in 1,932 FinnGen phenotypes and not classified as pathogenic/likely pathogenic in ClinVar (Kurki et al. 2022). ClinVar is a public and freely available archive of human genetic variants and interpretation of their relationships to diseases (Landrum et al 2020). GWAS is an observational study to identify genomic variants that are statistically associated with a disease or a specific trait. It is an efficient tool for identifying and observing associations of genotypes with phenotypes between individuals who are ancestrally similar but with different phenotypes and their differences in genetic allele variant frequency (Uffelmann et al. 2021). GWAS have identified rare (<0.5% allele frequency) and low-frequency (0.5-5%) variants in complex diseases such as prostate cancer (Gudmundsson et al. 2012).

Table 1. 51 coding variants identified by FinnGen-project (Kurki et al. 2022). Variants enriched in the Finnish population.
n = number of donors carrying variant in Biobank material
MAF = Minor allele frequency (Biobank)

rsID	n	gene	MAF-Biobank	consequence
rs147660927	2788	*ANGPTL7*	0.04020	missense
rs121909293	1483	*CTRC*	0.02117	missense
rs199935580	45	*THBS3*	0.00064	missense
rs35937944	1223	*COLGALT2*	0.01753	missense
rs11591147	2569	*PCSK9*	0.03691	missense
rs141266925	1015	*CASP7*	0.01496	missense
rs766868752	47	*SYNPO2L*	0.00066	splice donor
rs200852670	56	*SERPINB7*	0.00079	missense
rs142351376	1010	*IL4R*	0.01442	missense
rs144109867	1317	*TMEM119*	0.01877	stop gained
rs200317762	172	*TUBA1C*	0.00243	missense
rs148781286	367	*CD63*	0.00523	missense
rs201829738	214	*SMARCC2*	0.00302	missense
rs149722682	74	*LAG3*	0.00105	missense
rs201483470	503	*HHIPL1*	0.00718	missense
rs41531245	521	*LRRK1*	0.00739	missense
rs74653330	3045	*OCA2*	0.04404	missense
rs147301839	758	*MYZAP*	0.01078	missense
rs147972626	793	*RPL3L*	0.01130	missense
rs201864074	859	*RPL3L*	0.01223	missense
rs144651842	5135	*IL4R*	0.07532	missense
rs201162411	440	*CNGB1*	0.00624	missense
rs72553883	1700	*TNFRSF13B*	0.02428	missense
rs201955556	897	*GJD3*	0.01270	missense
rs138213197	503	*HOXB13*	0.00710	missense
rs199598395	658	*RNF43*	0.00935	missense
rs80338958	42	*SCN4A*	0.00059	missense
rs74006007	934	*CEP131*	0.01328	missense
rs201208667	445	*SERPINB7*	0.00631	missense
rs187429064	3474	*TM6SF2*	0.05059	missense
rs184042322	736	*AKT2*	0.01047	missense
rs150414818	609	*ALDH16A1*	0.00863	missense
rs371254530	21	*MYH14*	0.00030	missense
rs199600574	1231	*PPP1R26*	0.01753	missense
rs200336521	364	*NBEAL1*	0.00516	splice acceptor
rs77482050	3764	*ANO7*	0.05473	missense
rs145955907	1349	*ZAP70*	0.01924	missense
rs201557719	183	*PLTP*	0.00260	missense
rs780302457	51	*SLC35C2*	0.00072	missense
rs74203920	2405	*AIRE*	0.03456	missense
rs17879961	1975	*CHEK2*	0.02836	missense

rs45620037	351	*SCN5A*	0.00497	missense
rs776981958	70	*TERT*	0.00099	missense
rs116483731	2021	*SPDL1*	0.02888	missense
rs770636874	132	*RFX6*	0.00186	frameshift
rs745973283	37	*TNFAIP3*	0.00052	missense
rs62621812	2783	*ZNF800*	0.04010	missense
rs55960271	1174	*CLCN1*	0.01672	stop gained
rs771807370	76	*RFX6*	0.00107	frameshift
rs77273740	3200	*DBH*	0.04637	missense
rs199680517	1277	*PPP1R26*	0.01819	missense

1.5. Aims of the Pro graduate study and study questions

In the thesis project the aim is to evaluate frequencies and geographical distributions of genetic variants from the donor data stored in the Blood Service Biobank to understand the usefulness of blood donor samples and data in genomic studies. The aim is to compare the frequencies of the variants in the Biobank and FinnGen datasets and investigate how common disease associated variants are among blood donors. Another aim is to determine whether these variants are clustered to different areas of Finland.

The study questions are, can blood donor biobank also be valuable for identifying carriers with variants that are associated with diseases? Are there blood donors who have several disease associated variants and are there donors without any of these variants? How are these variants distributed in Finland based on the genetic population history.

2. Material and methods

2.1. Biobank material

The Blood Service Biobank samples are collected along blood donation from donors who have given a written broad biobank consent. Use of the samples and data is in accordance with the biobank consent and meets the requirements of the Finnish Biobank Act 688/2012. Whole blood donor samples were stored in +4 °C or if not extracted within 24 h from sampling, frozen to -20 °C. DNA was extracted from 35400 samples in biobank laboratory by using QiaSymphony magnetic bead technology for DNA purification and isolation (Qiagen QIAsymphony DSP DNA Midi Kit (937255)) and QiaSymphony DNA extraction instrument (Qiagen, 35306).

2.1.1. Blood donor DNA genotyping by FinnGen

The FinnGen-project aims to genotype 500 000 Finnish biobank participants. Genotyping, quality control, and genome imputation protocols are described in detail in FinnGen Gitbook (FinnGen 2022b). Biobank DNA samples were genotyped in the FinnGen-research project by customized Illumina and Affymetrix chip arrays (Illumina Inc., San Diego, and Thermo Fisher Scientific, Santa Clara, Ca, USA). The FinnGen- project carries out quality controls, pre-phasing and imputation for genotyped samples (FinnGen 2022c). Phasing is the process to infer haplotypes from known genotypes (Blackburn et al. 2020). For genotype imputation population specific SISU v3 reference panel was used. SISU reference panel consists of 3 775 whole genome sequenced data from Finnish individuals (FinnGen 2022e). Once the genotyping for blood donor extracted DNA has been carried out by FinnGen- project, the genomic data is returned to the Blood Service Biobank.

2.1.2. The Variant Call Format (VCF)

The genomic data is stored in Variant Call Format (VCF). Genetic DNA polymorphism data such as SNPs, insertion, deletions, and structural variants, can be stored as Variant Call Format (VCF) (Danecek et al. 2011). VCF file represents variants in which an individual's genome differs from a reference genome. A VCF file consists of two sections: a header section and data section. The header contains meta information and provides standardized description of the tags and annotations used in the data section. Currently used VCF files in Biobank are specification version 4.2. The VFC files from FinnGen contains following information: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT, AC, AN, NS, AF, MAF, AC_Het, AC_Hom, AC_Hemi, HWE and ExcHet (table 1.). In FinnGen VCF file FORMAT field has GT:DS:GP, where GT means genotype, DS means estimated alternate allele dosage, calculated as [P(0/1)+2*P(1/1)] and GP means estimated Posterior Genotype Probabilities P(0|0), P(0|1) and P(1|1). Genotype 0|0 = sample is homozygous reference, 0|1 = sample is heterozygous and carrying 1 copy of each of REF and ALT alleles, 1|1 = sample is homozygous alternate and .|. = no genotype called, or the genotype is missing. Estimated posterior probabilities have three subfields for homozygous reference, heterozygous and homozygous alternate genotypes probabilities. For example, FORMAT field GT:DS:GP 0|0:0.01:0.99,0.01,0 gives the information that the sample is homozygous (0|0): the expected probability of the alternate allele is 0.01: genotype posterior probabilities are 0.99,0.01,0 where genotype is homozygous as the highest probability is 0.99

Table 2. VCF file information.

#CHROM	Chromosome
POS	Position
ID	Identification
REF	Reference base(s)
ALT	Alternative allele
QUAL	Quality
FILTER	Filter status
INFO	Properties observed at the level of the variant
FORMAT	Contains sample specific information such as genotype and individual sample-level annotation values

AC	Allele count in genotypes, for each ALT allele
AN	Total number of alleles in called genotypes
NS	Number of samples with data
AF	Allele frequency
MAF	Minor Allele frequency
AC_Het	Allele counts in heterozygous genotypes
AC_Hom	Allele counts in homozygous genotypes
AC_Hemi	Allele counts in hemizygous genotypes
HWE	Hardy-Weinberg equilibrium test
ExtHet	Probability of excess heterozygosity

2.2.Data filtering

First the genomic location of the 51 identified variants (table 1.) was checked by the marker rs identification number on Ensembl database in human genome build 38. Based on the genomic position, genome data was manipulated on Linux and after that transferred to RStudio. Manipulation and variant calls were done using BCFtools commands on Blood Service Biobank's own Linux server, where the genomic data is maintained and stored. BCFtools is a variant call data manipulation utility set (Samtools 2022). BCFtool commands were run on integrated development environment, RStudio, using R programming language. With BCFtools the 51 variants were filtered from the genome data as allelic letter form (A,T,C,G) and dosage form (0|0, 0|1/1|0, 1|1) based on the donor's genotype.

2.3. Data identification

First, all the major allele heterozygous with allele combination 0|0, minor allele heterozygous with allele combinations 0|1 or 1|0 and minor allele homozygous 1|1 donors for each of the 51 variants were identified and counted (see appendix 1). The total number of each variant found as heterozygous or homozygous among the blood donor data was examined. By filtering only major allele homozygous 0|0 allele combinations from the genetic data, we were able to recognize the donors without any of the investigated markers.

Differences between hetero- and homozygosity were explored by comparing the variants together. The minor allele frequency (MAF) was counted from Biobank material. Major allele is the most common allele, reference allele, and minor allele is the second most common allele. MAF is calculated as following: number of copies of allele in population/total number of gene in population.

2.4. Finnish province division.

Division between provinces was done based on Kerminen et al. (2017) research, where they had divided Finnish population into East and West based on fineSTRUCTURE results and based on strong genetic difference between western and eastern regions (Lappalainen et al. 2006). Donors' postal codes were combined to the genetic data. The postal codes of the donors were divided into 19 provinces in Finland based on Statistics Finland province map (Tilastokeskus 2021). Eastern provinces are Kainuu (KAI), North Karelia (PKAR), South Savo (ESAV), South Karelia (EKAR), North Savo (PSAV), North Ostrobothnia (POH), Central Finland (KSU), Päijät-Häme (PÄI), Kymenlaakso (KYM). Western provinces are Uusimaa (UUS), Tavastia Proper (KHÄ), Pirkanmaa (PIR), Southwest Finland (VAR), Satakunta (SAT), Åland (AHV), South Ostrobothnia (EPO), Ostrobothnia (POH), Central Ostrobothnia (KPO), Lapland (LAP). The western and eastern origin of the donors was judged based on their postal codes (red line in fig. 1) (see appendix 7).

Figure 1. 19 provinces in Finland, western and eastern provinces are divided with red line (Tilastokeskus 2019).

The total number of blood donors in each of the 19 provinces was calculated based on the blood donors' postal code information. Donor postal codes were allocated to the correct provinces. Minor allele frequencies of the 51 rare diseases were calculated for each of the 19 provinces together with east/west areas (see appendix 6).

2.5.Statistical analysis

2.5.1. Spearman's correlation test

Spearman's rank correlation was used to compare the minor allele frequencies (MAF) between the Blood Service Biobank and FinnGen, gnomAD-FIN and gnomAD-NFSEE (Genome Aggregation Database) datasets (see appendix 3). GnomAD-FIN minor allele frequencies were used as a reference for the Finnish population and gnomAD-NFSEE minor allele frequencies as a European reference excluding the Finnish-Swedish-Estonian population.

2.5.2. Chi-squared test

Pearson's chi-squared (χ^2) test was used to compare distributions of the variants between East and West of Finland (see appendix 6). Chi-square test was used to determine if the observed proportion of hetero- and homozygous variant carriers differ between East and West parts of Finland (fig. 1). Test was performed for each of the 42 variants that were detected in more than 30 donors. Chi-squared test relies on an approximation assuming a large sample.

2.5.3. Fisher's exact test

Fisher's exact test was used to determine if the incidence of observed variants is different between East and West part of Finland with the 9 test groups with less than 30 carriers, by comparing the number of donors carrying the specific variant and the overall number of donors in a contingency table (see appendix 6). Fisher's exact test was also used to examine if observed variants are significantly clustered in specific provinces based on the visual comparison of minor allele frequency of the 8 variants in each province (fig. 8).

Fisher's exact test is usually used for small sample size, when there is one or more small cells in the contingency table, in this study when $n < 30$ (Kim 2017).

2.5.4. Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium exact test was performed for each of the 51 variants independently. For the allele frequencies p and q (p + q = 1), the Hardy-Weinberg equilibrium (HWE) gives the possible frequencies for genotypes AA, AB and BB with respected frequencies of p2 + 2pq + q2 = 1. (Mayo 200).

First, the number of donors carrying all of the three possible genotypes (AA, AB, BB) were counted for each variant. The HWE test was done by using HardyWeinberg R package on RStudio, which performs the exact test for Hardy-Weinberg equilibrium (Zhao 2007) (see appendix 4).

2.5.5. Multiple P-value adjustment

All of p-values from statistical analyses (Spearman's correlation, chi-squared test, Fisher's exact test and Hardy-Weinberg exact test) used in this project, were adjusted using Benjamini & Yekutieli "BY" adjustment method for controlling the False Discovery Rate (FDR) (Benjamini & Yekutieli 2001) (see appendix 5). The p-value adjustment method controls the FDR at the wanted level q for independent and positively dependent test statistics and is necessary for multiplicity testing. False discovery rate means expected portion of discoveries which are falsely rejected (Benjamini & Hockerberg 1995). Multiple p-value adjustment is needed because multiple tests were run simultaneously increasing the probability that one or more null hypotheses will be rejected incorrectly. Q-values are adjusted p-values and will be used for interpreting the final results in the project. By using q-values the risk of getting false statistically significant results caused by multiple testing and number of differences will decrease.

2.6. Principal Component analysis

The results of whole genome Principal Component analysis (PCA) for the population structure were received from the FinnGen-project. FinnGen-project has calculated 20 principal component points for each of the donors (FinnGen 2022d). Principal component analysis (PCA) is usually performed to create dimensional reduction to ease visualization in large datasets while retaining most of the information and minimizing information lost. This reduction is accomplished by identifying principal components which explain variance in the data (Ringner 2008). With principal components the data can be presented in such a way that only the largest part of the variation is being used, usually the first few components. The idea is to preserve as much variability as possible, meaning that finding new variables that are linear combinations of those in the original data. Those variables successively preserve the maximum amount of variance and are uncorrelated with each other. This is how most of the information is retained but dimensionality is reduced. (Jolliffe & Cadima 2016).

Principal component analysis gives information about how much variance is explained by each of the principal component and how much data can be reduced. The less there are components that explain the overall variation in the data, the more overlapping variation the variables have. The more independent the variables, the less variation a single component explains. Variables explained by the components show the variation similarity of the variables. If only a few of the principal components explain most of the variation in the data, variables are similar. The first principal component explains most of the variation in the dataset, the second component the second most variance and so on.

2.6.1. Explained variation by principal components

Twenty principal components calculated by FinnGen-project for each of the 35 400 donors were returned back to Biobank. The percentage of the explained variance had to be calculated for each of the principal components (see appendix 2). By counting the variance of each one of the PC components divided by the sum of the variance of the other PC components, we were able to see how much of the variation each principal component explains in the data (fig. 2). PC1 explains 32% of the variation and PC2 explains 11% of the variation and so on. All of the 20 principal components together explain 100% of the data. Based on the explained variance, PC1 and PC2 were chosen for further investigation in this study. By plotting each of the donors' PC1 and PC2, we were able to investigate the possible difference on whole genome level between those donors carrying a certain variant and those who don't carry the same variant.

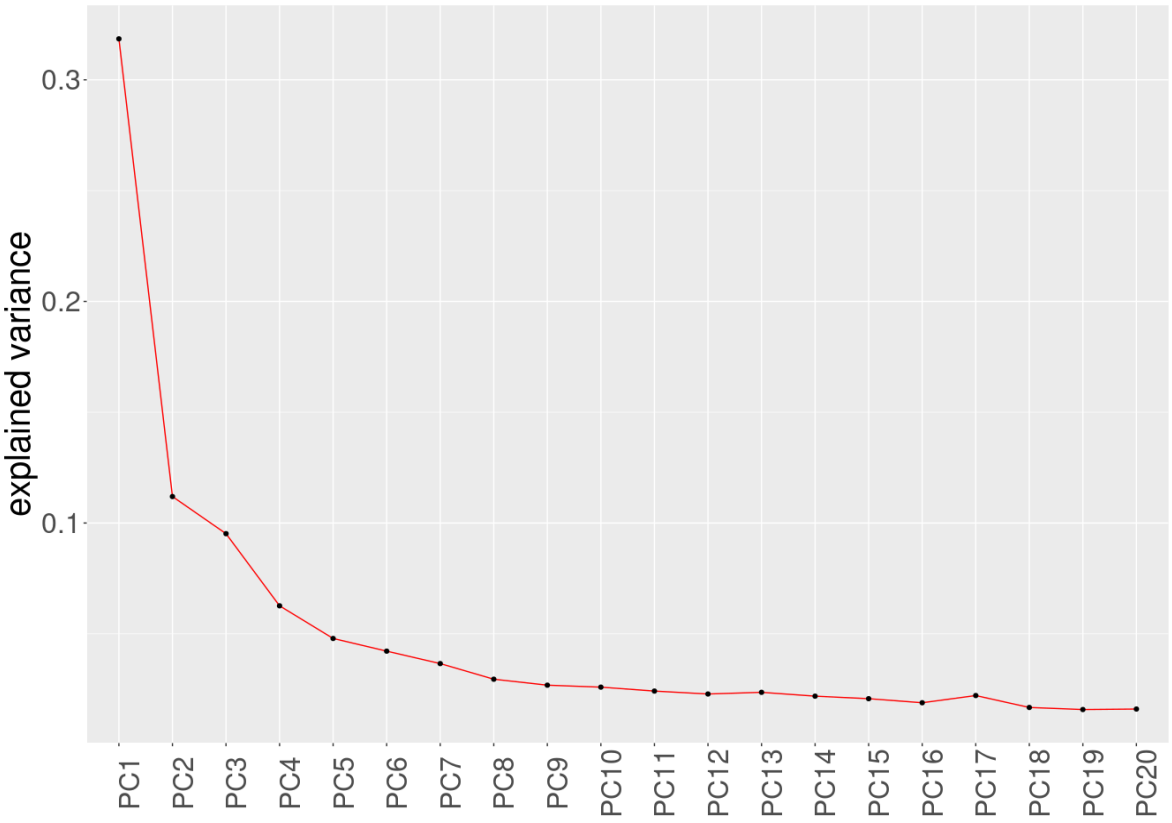


Figure 2. The proportion of explained variance counted for each principal component.

2.6.2. Variation in Finnish provinces

For investigating how principal component points have been distributed between 19 Finnish provinces, the median for the first 10 principal components was counted for each of the 19 provinces to understand central tendency of a set of statistical scores. Median is more suitable for skewed distribution to derive central tendency because it is more robust and sensible than mean (Kenney & Keeping 1962). Blood donation data also contains donor postal code information, and the postal code information was combined with the donors PCA results. All the donors without PCA result and postal code information were removed and after that separated into 19 provinces based on the donor's postal code how the data is distributed across the provinces. Median for 10 principal components was calculated and plotted on the map of Finland to see if there are any differences in donors' principal components between provinces. Map of Finland R package (Haukka 2022) was used for the analysis (see appendix 2). With the R Package, we were able to map the provinces on the map of Finland based on their geometric values.

All analyses in this study were performed with R version 3.6.1 and 4.0.5 (R Core Team 2021) with Rstudio (RStudio Team 2021) (see appendix 1-6).

3. Results

Of the entire study set Biobank donors (n=35400) 148 had no postal code and were excluded from the analysis. Altogether 35252 donors in total were divided in to 19 Finnish provinces based on donor's postal code (Fig. 3). Most of the donors are residents of Uusimaa and Southewest Finland provinces. There are less donors (n=10585) in Eastern provinces of Finland than Western provinces (n=24667).

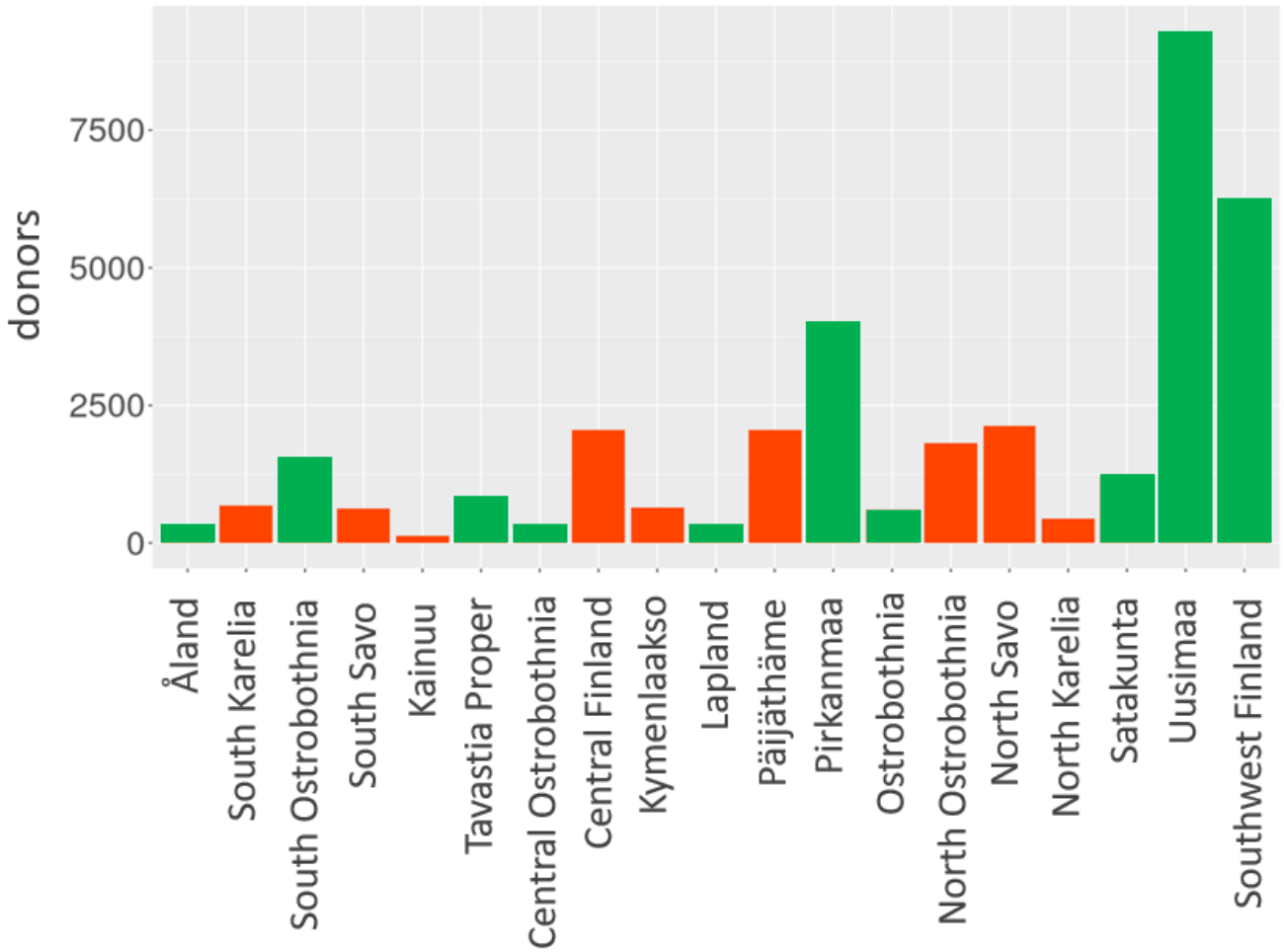


Figure 3. Total number of donors (n=35252) with postal code information in 19 Finnish provinces. Western provinces are colored green and Southern provinces are colored orange.

3.1. Prevalence of variants among blood donors

All of the 51 rare disease associated variants enriched in Finland (table 1.) were detected from blood donors (n=35400). Some donors carried more than one variant, whereas 20,4 % did not carry any of the 51 variants (n=7229 donors in total). None of the donors carried all of the 51 variants. While all variants were found heterozygous and 14 of the variants were not found homozygous at all. The Most common variant is rs144651842, which causes *IL4R* gene missense mutation is associated with chronic lower respiratory disease (table 1). In this study cohort, altogether 5135 carriers (14,5 %) were found for this variant, of which 4937 are heterozygous and 198 homozygous for the variant. The most uncommon variant with the least carriers is rs371254530 that is associated with sensorineural hearing loss. This variant has 21 heterozygous carriers (table 3.). Figures 4 and 5 show the variant prevalence between heterozygous and homozygous donors in biobank material. Donors carrying minor allele homozygous variant are not as common as heterozygous carriers because homozygote genotype needs to inherit the disease associated variant from both of the parents.

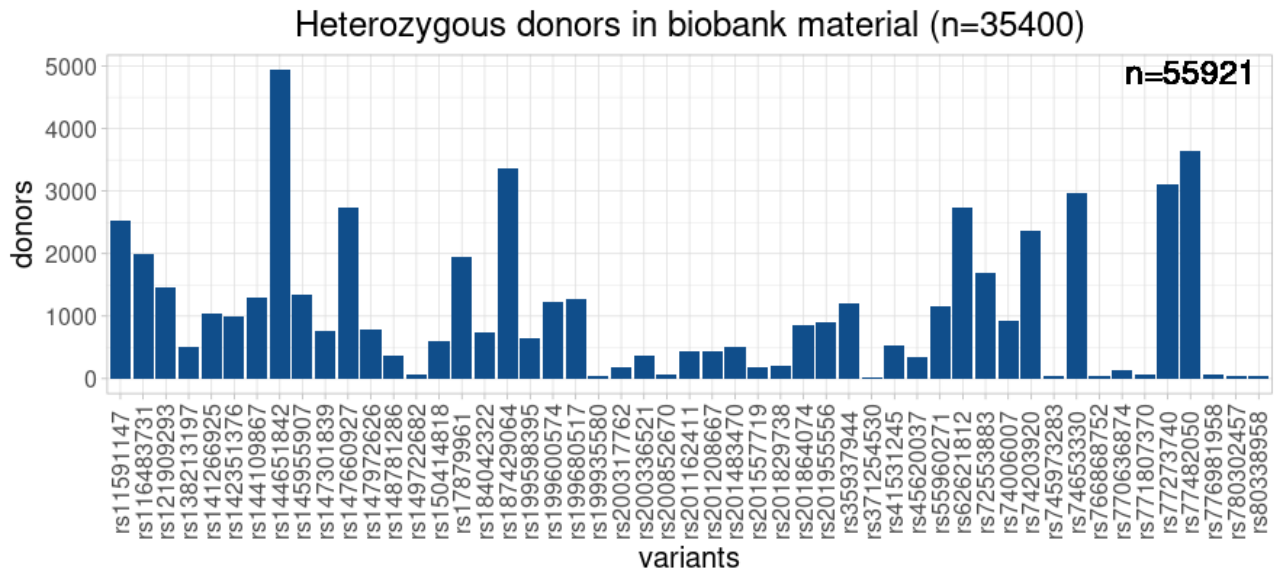


Figure 4. Variant heterozygotes donors in FRC Blood Service Biobank. 51 variants identified by FinnGen.

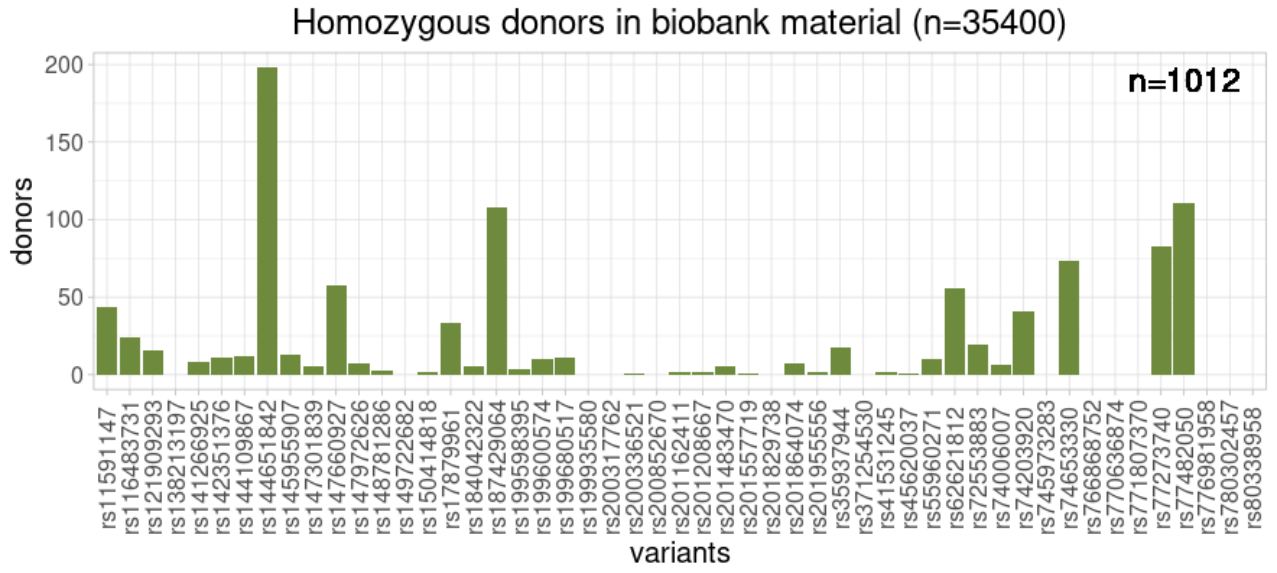


Figure 5. Variant homozygotes in FRC Blood Service Biobank. 51 variants identified by FinnGen.

Table 3. Number of donors carrying each variant in biobank material

rsID	major allele homozygotes	minor allele heterozygotes	minor allele homozygotes	in total	frequency
rs371254530	35379	21	0	21	0.0297
rs11591147	35363	37	0	37	0.0523
rs80338958	35358	42	0	42	0.0593
rs199935580	35355	45	0	45	0.0636
rs766868752	35353	47	0	47	0.0664
rs780302457	35349	51	0	51	0.0720
rs200852670	35344	56	0	56	0.0791
rs121909293	35330	70	0	70	0.0989
rs149722682	35326	74	0	74	0.1045
rs200852670	35324	76	0	76	0.1073
rs35937944	35268	132	0	132	0.1864
rs200317762	35228	172	0	172	0.2429
rs201557719	35217	182	1	183	0.2599
rs201829738	35186	214	0	214	0.3023
rs147660927	35049	350	1	351	0.4972
rs200336521	35036	363	1	364	0.5155
rs148781286	35033	364	3	367	0.5226
rs201162411	34960	438	2	440	0.6243
rs201208667	34955	443	2	445	0.6314
rs138213197	34897	503	0	503	0.7105
rs201483470	34897	498	5	503	0.7175
rs41531245	34879	519	2	521	0.7387
rs150414818	34791	607	2	609	0.8630
rs199598395	34742	654	4	658	0.9350
rs184042322	34664	731	5	736	1.0466
rs147301839	34642	753	5	758	1.0777
rs147972626	34607	786	7	793	1.1299
rs201864074	34541	852	7	859	1.2232
rs201955556	34503	895	2	897	1.2698
rs74006007	34466	928	6	934	1.3277
rs142351376	34390	999	11	1010	1.4421
rs141266925	34349	1043	8	1051	1.4958
rs766868752	34226	1164	10	1174	1.6723
rs35937944	34177	1205	18	1223	1.7528
rs199600574	34169	1221	10	1231	1.7528
rs144109867	34123	1266	11	1277	1.8192
rs144109867	34083	1305	12	1317	1.8771
rs145955907	34051	1336	13	1349	1.9237
rs121909293	33917	1467	16	1483	2.1172
rs72553883	33700	1681	19	1700	2.4280

rs17879961	33425	1942	33	1975	2.8362
rs199935580	33379	1997	24	2021	2.8884
rs74203920	32994	2365	41	2406	3.4562
rs11591147	32831	2525	44	2569	3.6907
rs141266925	32617	2727	56	2783	4.0099
rs147660927	32612	2730	58	2788	4.0198
rs74653330	32355	2972	73	3045	4.4040
rs142351376	32200	3117	83	3200	4.6370
rs187429064	31926	3366	108	3474	5.0593
rs77482050	31636	3653	111	3764	5.4732
rs144651842	30265	4937	198	5135	7.5325

3.2. Minor allele frequency correlation

The minor allele frequency comparison between FRC Biobank blood donor population and FinnGen hospital population show how allele frequencies of these variants are distributed between "healthy" and diseased populations. Minor allele frequencies of FRC Biobank dataset were compared with FinnGen dataset, and another Finnish reference dataset, gnomAD-FIN, and also with non-Finnish-Swedish-Estonian reference dataset, gnomAD-NFSEE (Fig. 6). Results show a strong correlation between Biobank and FinnGen variant frequencies based on Spearman's correlation result (p-value = 4.89E-47, q-value = 2.93E-44; rho = 0.9930). Comparison between Biobank and gnomAD-FIN frequencies shows strong correlation as well (p-value = 6.21E-40, q-value = 1.86E-37; rho 0.9862). Variant frequency comparison between Biobank and gnomAD-NFSEE show less strong correlation (p-value = 9.56E-11, q-value = 6.38E-09; rho 0.7605) as most of the variants fall outside the 95% confidence coefficient.

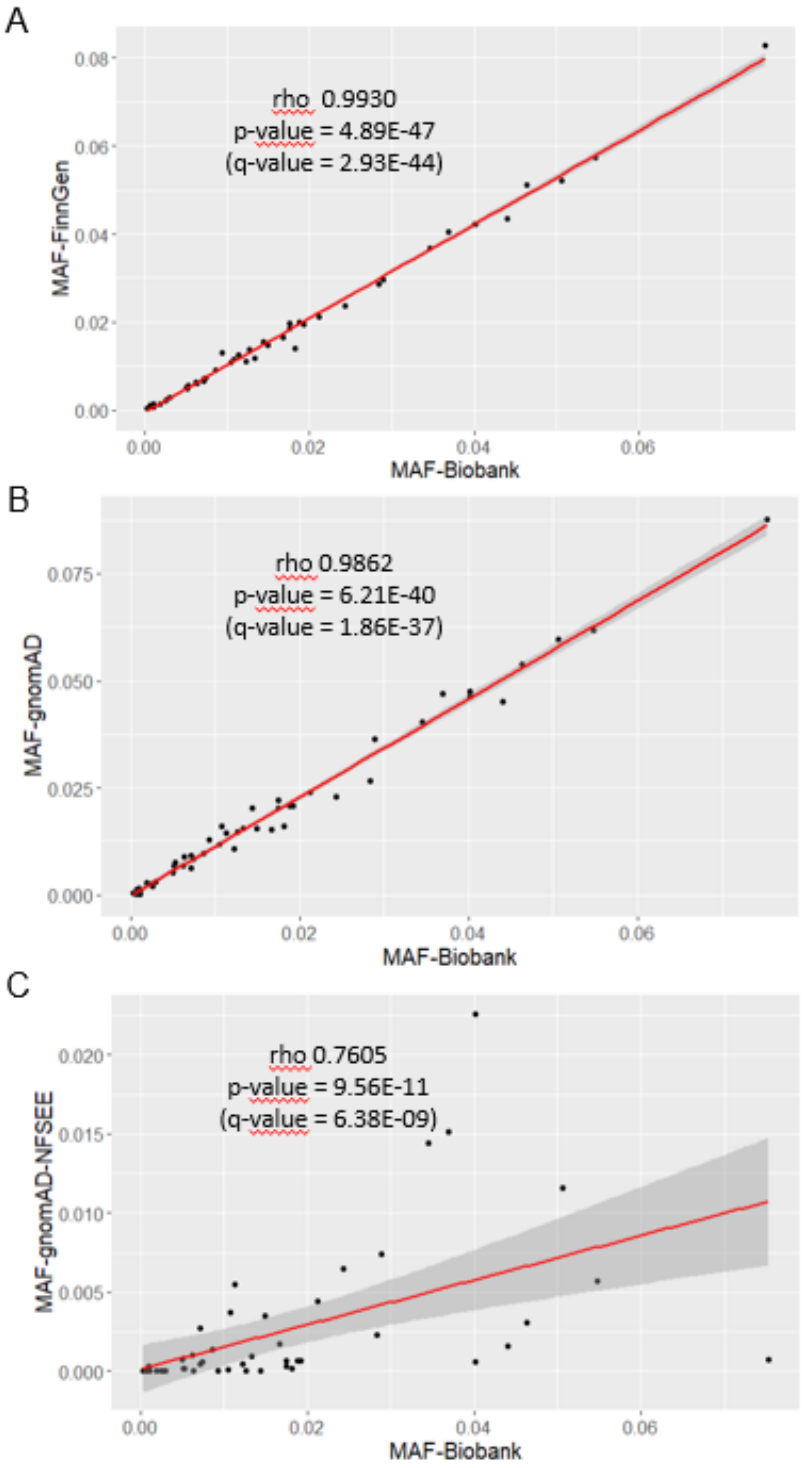


Figure 6. Minor allele frequency correlation between Biobank & FinnGen, Biobank & gnomAD-FIN, Biobank & gnomAD-NFSEE.

A. Spearman's rank correlation; p-value = 4.89E-47, q-value = 2.93E-44, rho 0.9930
B. Spearman's rank correlation; p-value = 6.21E-40, q-value = 1.86E-37, rho 0.9862
C. Spearman's rank correlation; p-value = 9.56E-11, q-value = 6.38E-09, rho 0.7605

3.3. Variant frequency comparison between Eastern and Western provinces in Finland

Nineteen Finnish provinces were divided into two groups, East and West, based on their geographical location. Of all of the 51 variants, 29 variants had higher frequency in the East and 12 variants had higher frequency in the West. Difference in the prevalence of variants with more than 30 carriers between East and West were counted for 42 variants in contingency table using Chi-squared method. 30 variants showed statistically significant difference based on chi-squared test results based on p-value < 0.05, twenty four of them showed statistically significant difference based on q-values. Six results were discarded as false positive, where q-values showed no significant difference. Twenty four variants, that had significant difference, had higher frequency in the East based on the p-value and 22 variants had higher frequency in the East based on the q-value. Six variants had higher frequency in the West based on the p-value and two of them had higher frequency based on the q-value. Twelve variants didn't show statistically significant difference between East and West based on both p- and q-values (table 4). P-values were adjusted using Benjamini & Yekutieli (1995) adjustment method "BY".

Table 4. Chi-squared test for 42 variants, with < 30 carries between donors in East and West. Statistically significant differences between East and West are highlighted with green color.

chi-squared test	rs147660927	rs121909293	rs35937944	rs11591147	rs141266925	rs142351376	rs144109867	rs200317762
p-value	3.52E-10	6.35E-04	1.48E-07	1.08E-06	7.84E-06	4.85E-16	1.32E-03	1.81E-03
q-value	1.98E-08	1.36E-02	5.72E-06	3.52E-05	2.11E-04	7.50E-14	2.55E-02	3.11E-02
donors carrying variant in East (n)	978	504	449	876	381	418	445	70
donors carryin variant in West (n)	1795	978	771	1679	670	587	863	101
donors in East (n)	9607	10081	10136	9709	10204	10167	10140	10515
donors in West (n)	22872	23689	23896	22988	23997	24080	23804	24566
chi-squred test	rs148781286	rs201829738	rs201483470	rs41531245	rs74653330	rs147301839	rs147972626	rs201864074
p-value	4.06E-05	3.83E-04	1.04E-01	1.82E-01	3.18E-07	6.45E-12	2.79E-01	6.76E-04
q-value	1.05E-03	9.11E-03	1.00E+00	1.00E+00	1.16E-05	5.70E-10	1.00E+00	1.39E-02
donors carrying variant in East (n)	145	88	167	170	1036	313	251	212
donors carryin variant in West (n)	219	126	334	350	2003	444	539	644
donors in East (n)	10440	10497	10418	10415	9549	10272	10334	10373
donors in West (n)	24448	24541	24333	24317	22664	24223	24128	24023
chi-squred test	rs144651842	rs201162411	rs72553883	rs201955556	rs138213197	rs199598395	rs74006007	rs201208667
p-value	3.18E-09	1.69E-01	2.93E-01	1.94E-09	4.58E-01	2.08E-01	1.51E-03	5.80E-01
q-value	1.40E-07	1.00E+00	1.00E+00	9.24E-08	1.00E+00	1.00E+00	2.84E-02	1.00E+00
donors carrying variant in East (n)	1716	144	489	350	158	211	323	128
donors carryin variant in West (n)	3401	292	1204	545	3.43E+02	443	607	316
donors in East (n)	8869	10441	10096	10235	10427	10374	10262	10457
donors in West (n)	21266	24375	23463	24122	24324	24224	24060	24351
chi-squred test	rs187429064	rs184042322	rs150414818	rs199600574	rs200336521	rs77482050	rs145955907	rs201557719
p-value	6.52E-13	2.91E-01	2.74E-02	1.42E-11	4.29E-03	7.51E-10	2.20E-02	1.66E-03
q-value	8.07E-11	1.00E+00	3.85E-01	1.10E-09	6.99E-02	3.87E-08	3.16E-01	2.93E-02
donors carrying variant in East (n)	1223	234	157	475	147	1289	441	35
donors carryin variant in West (n)	2237	502	448	752	448	2460	902	146
donors in East (n)	9362	10351	10428	10110	10438	9296	10144	10550
donors in West (n)	22430	24165	24219	23915	24219	22207	23765	24521
chi-squred test	rs74203920	rs17879961	rs45620037	rs116483731	rs770636874	rs62621812	rs55960271	rs771807370
p-value	1.87E-12	6.03E-01	3.27E-03	5.47E-08	9.45E-01	1.53E-04	5.46E-01	1.08E-02
q-value	1.93E-10	1.00E+00	5.47E-02	2.26E-06	1.00E+00	3.78E-03	1.00E+00	1.63E-01
donors carrying variant in East (n)	874	600	80	712	40	921	342	33
donors carryin variant in West (n)	1528	1364	270	1298	92	1854	828	43
donors in East (n)	9711	9985	10505	9873	10545	9664	10243	10552
donors in West (n)	23139	23303	24397	23369	24575	22813	23839	24624
chi-squred test	rs77273740	rs199680517						
p-value	4.12E-04	3.23E-02						
q-value	9.44E-03	4.44E-01						
donors carrying variant in East (n)	1041	347						
donors carryin variant in West (n)	2136	923						
donors in East (n)	9544	10238						
donors in West (n)	22531	23744						

Fisher's exact test was performed for 9 variants with less than 30 carriers in the contingency table. Three variants showed statistically significant difference between East and West based on p-value and only one variant showed significant difference based on q-value (table 5.).

Table 5. Fisher's exact test for 9 variants between donors in East and West. Statistically significant difference between East and West is highlighted with green color.

Fisher's exact test	rs199935580	rs766868752	rs200852670	rs149722682	rs80338958
p-value	1.58E-03	6.47E-03	1.90E-01	1.28E-01	4.98E-01
q-value	2.88E-02	1.03E-01	1.00E+00	1.00E+00	1.00E+00
donors carrying variant in East (n)	24	23	12	16	10
donors carryin variant in West (n)	21	24	44	58	31
donors in East (n)	10561	10562	10573	10569	10575
donors in West (n)	24646	24643	24623	24609	24636
Fisher's exact test	**rs371254530**	**rs780302457**	**rs776981958**	**rs745973283**	
p-value	2.33E-01	8.08E-03	6.03E-01	2.08E-01	
q-value	1.00E+00	1.25E-01	1.00E+00	1.00E+00	
donors carrying variant in East (n)	9	24	23	15	
donors carryin variant in West (n)	12	26	47	22	
donors in East (n)	10576	10561	10562	10570	
donors in West (n)	24655	24641	24620	24645	

The four statistically most significant variants based on the chi-squared test results (table 4.) also visually show differences in minor allele frequencies between East and West (fig. 7.). Three out of the four variants show higher frequency in the East and one shows higher frequency in the West.

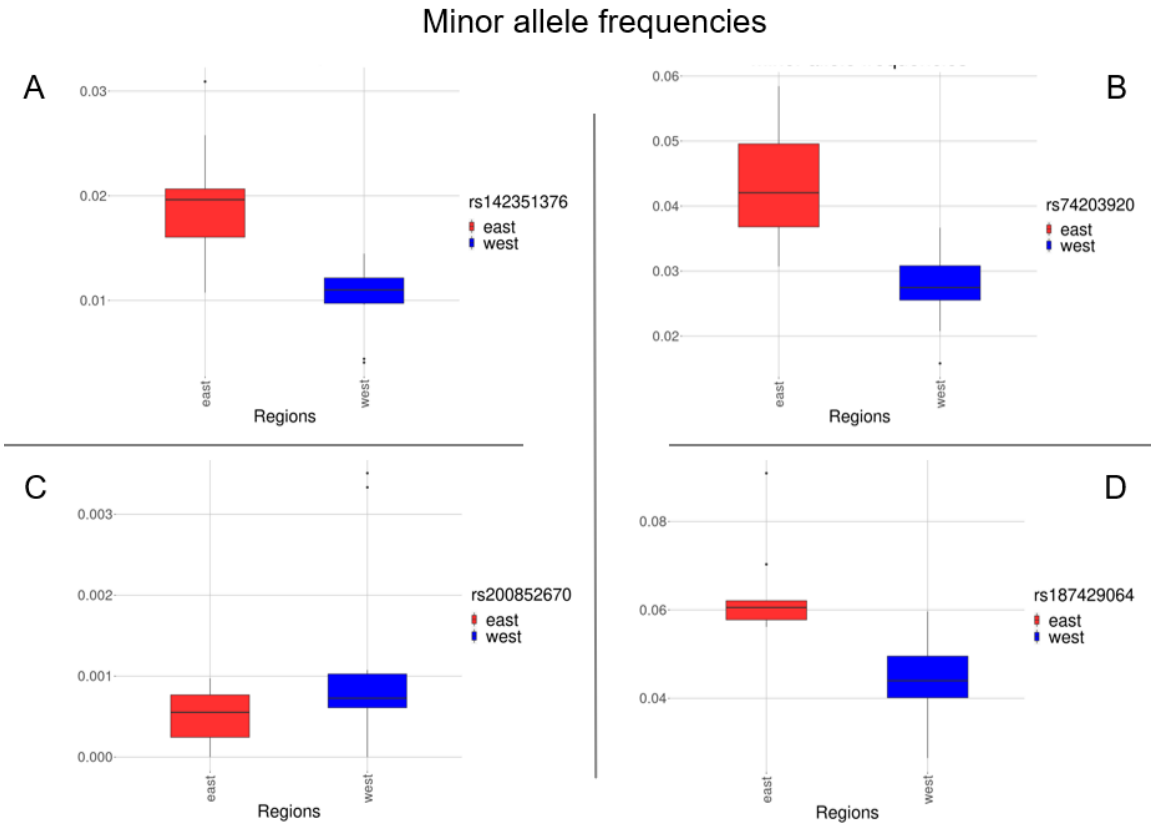
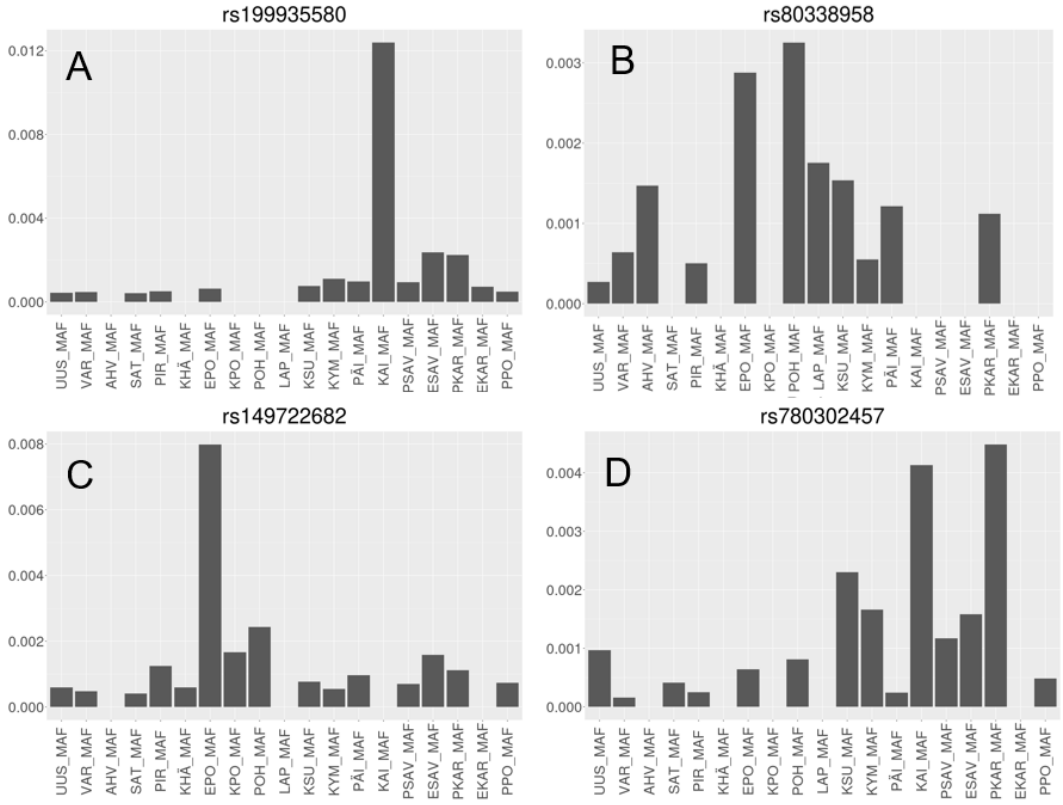


Figure 7. The four variants, which differ statistically most significantly in terms of prevalence between Eastern and Western provinces

3.4.Variant distribution between the Finnish provinces

All variants' minor allele frequencies in each of the 19 provinces were visualized using bar charts to see if there is visually significant difference between provinces. Eight variants seemed provisionally different in their frequency between provinces (fig. 8). Fisher's exact test was used to examine whether there is statistically significant difference between a specific province and rest of Finland. Six of the variants showed significant difference (p-value < 0.05) between the specific province and all other provinces. Based on q-values, 5 variants showed significant difference (table 5). One variant was discarded as false positive. The leading difference was observed with the rs149722682 between South Ostrobothnia and rest of the Finland.



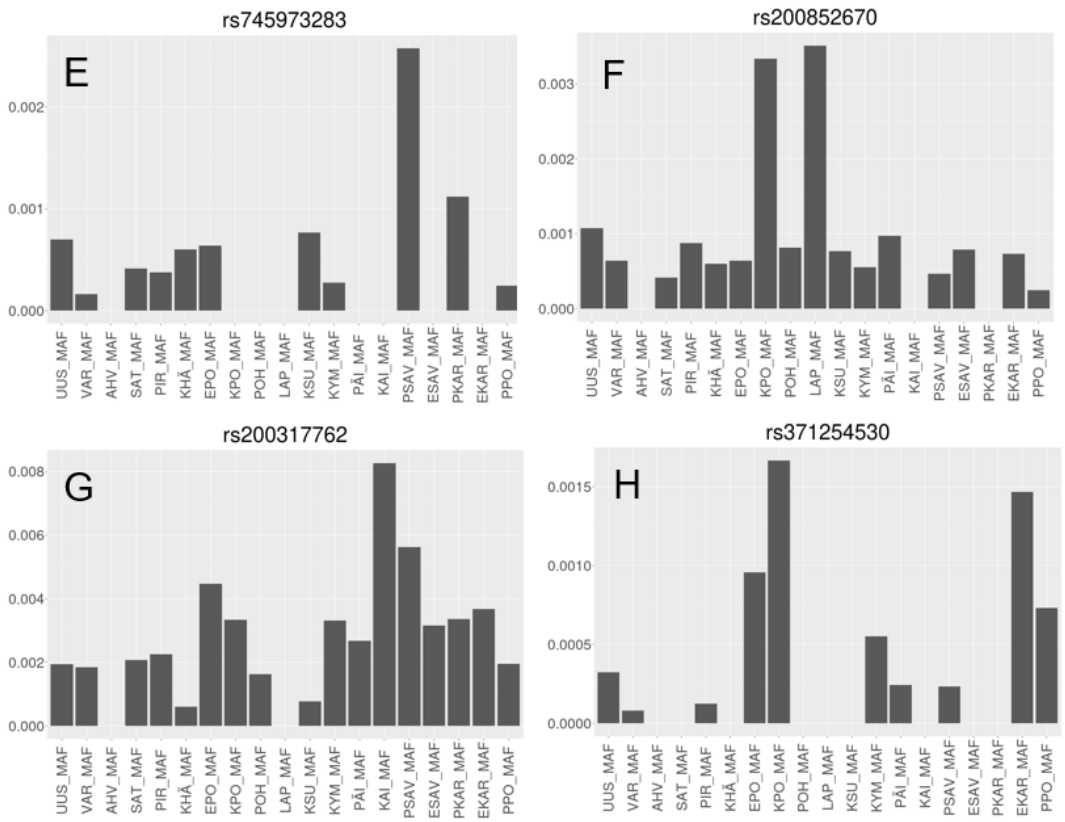


Figure 8. Variant frequencies in the 19 Finnish provinces. Statistically significant difference in variant frequencies based Fisher's exact test

A. rs199935580; Kainuu vs rest if the Finland (p-value = 4.99E-04, q-value = 1.10E-02).
B. rs80338958; South Ostrobothnia and Ostrobothnia vs rests of the Finland (p-value = 4.43E-07, q-value = 1.52E-05).
C. rs149722682; South Ostrobothnia vs rest of the Finland (p-value = 2.20E-16, q-value = 4.54E-14).
D. rs780302457; Kainuu and North Karelia vs rest of the Finland (p-value = 1.26E-03, q-value = 2.51E-02).
E. rs745973283; North Savo vs rest of the Finland (p-value = 4.40E-06, q-value = 1.24E-04)
F. rs200852670; Central Ostrobothnia and Lapland vs rest of the Finland (p-value = 1.32E-02, q-value = 1.95E-01).
G. rs200317762; Kainuu and rest of the Finland (p-value = 1.18E-01, q-value = 1)
H. rs371254530; Central Ostrobothnia vs rest of the Finland (p-value = 1.63E-01, q-value = 1).

Table 5. Variants compared between specific province and rest of the provinces in Finland using Fisher's exact test. Statistically significant differences are highlighted with green.

Fisher's exact test	rs199935580	Fisher's exact test	rs371254530
p-value	4.99E-04	p-value	1.63E-01
q-value	1.10E-02	q-value	1.00E+00
donors carrying variant in Kainuu (n)	3	donors carrying variant in Keski-Pohjanmaa (n)	1
donors carryin variant in Finland (n)	42	donors carryin variant in Finland (n)	20
donors in Kainuu (n)	118	donors in Keski-Pohjanmaa (n)	299
donors in Finland (n)	35210	donors in Finland (n)	35232
Fisher's exact test	**rs200317762**	**Fisher's exact test**	**rs745973283**
p-value	1.18E-01	p-value	4.40E-06
q-value	1.00E+00	q-value	1.24E-04
donors carrying variant in Kainuu (n)	2	donors carrying variant in Pohjois-Savo (n)	11
donors carryin variant in Finland (n)	170	donors carryin variant in Finland (n)	26
donors in Kainuu (n)	119	donors in Pohjois-Savo (n)	2124
donors in Finland (n)	35082	donors in Finland (n)	35226
Fisher's exact test	**rs80338958**	**Fisher's exact test**	**rs200852670**
p-value	4.43E-07	p-value	1.32E-02
q-value	1.52E-05	q-value	1.95E-01
donors carrying variant in Etelä-Pohjanmaa and Pohjanmaa (n)	13	donors carrying variant in Keski-Pohjnamaa and Lappi (n)	4
donors carryin variant in Finland (n)	29	donors carryin variant in Finland (n)	52
donors in Etelä-Pohjanmaa and Pohjanmaa (n)	2167	donors in Keski-Pohjanmaa and Lappi (n)	581
donors in Finland (n)	35223	donors in Finland (n)	35200
Fisher's exact test	**rs149722682**	**Fisher's exact test**	**rs780302457**
p-value	2.20E-16	p-value	1.26E-03
q-value	4.54E-14	q-value	2.51E-02
donors carrying variant in Etelä-Pohjanmaa (n)	25	donors carrying variant in Kainuu ad Pohjois-Karjala (n)	5
donors carryin variant in Finland (n)	49	donors carryin variant in Finland (n)	46
donors in Etelä-Pohjanmaa (n)	1540	donors in Kainuu and Pohjois-Karjala (n)	562
donors in Finland (n)	35203	donors in Finland (n)	35206

3.5. Principal Component Analysis in Finnish provinces

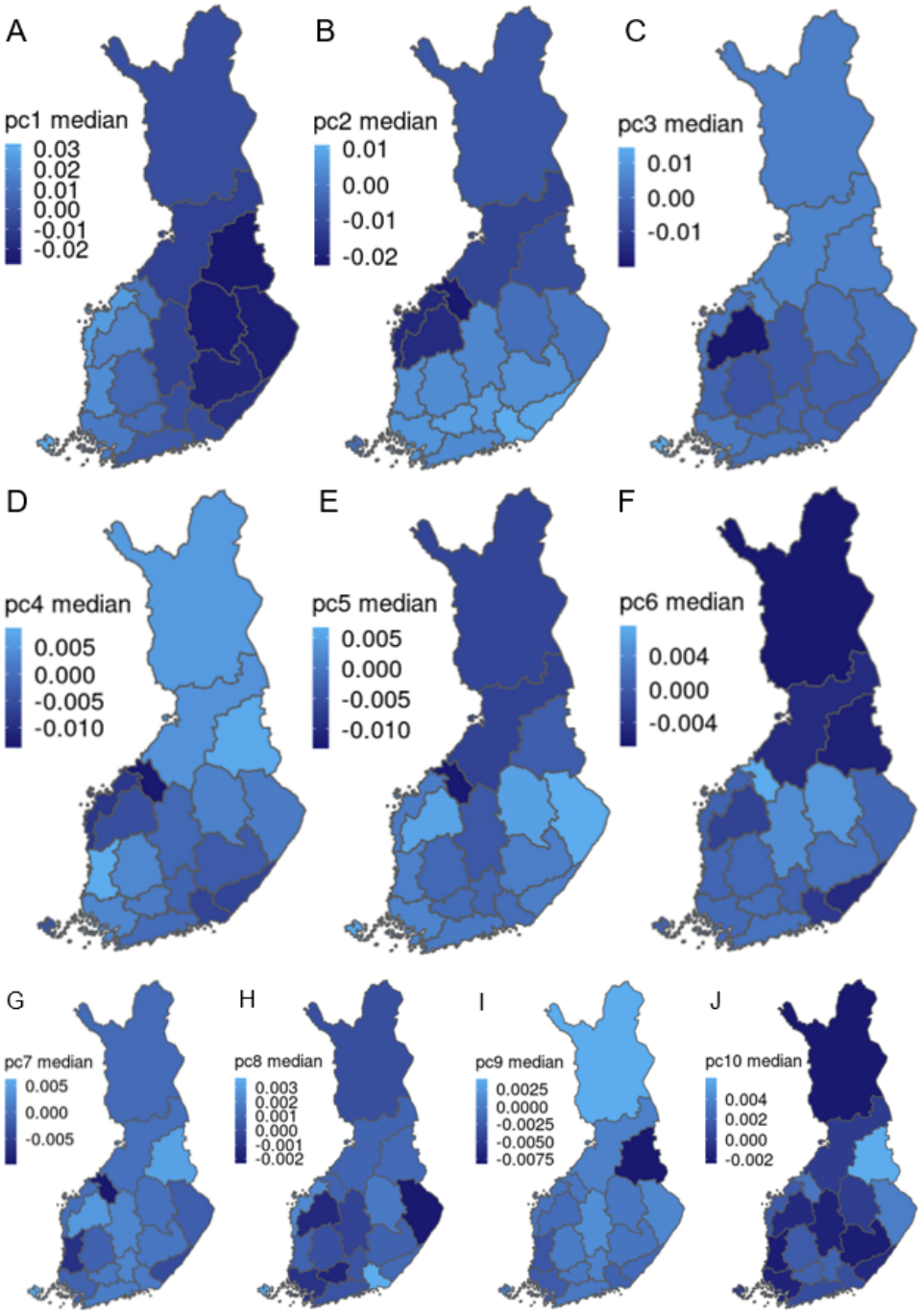
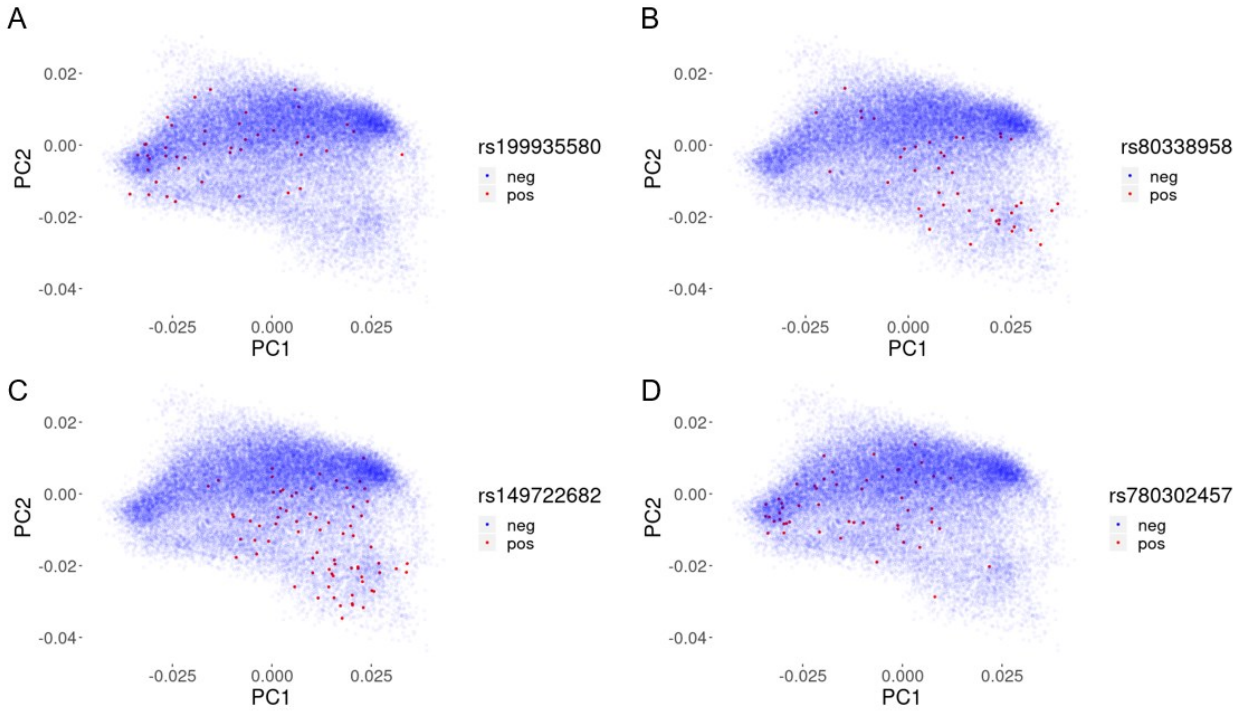


Figure 9. Distribution of 10 principal components between Finnish provinces.

PC results were divided between provinces based on blood donor postal code information. Each of the donors was placed in the province where they have stated they live. Median for 10 principal components show difference in PC results in each of the 19 provinces. PC1 explains most of the variance and shows clear difference between East and West. PC2 explains most of the variation in Ostrobothnia and PC3 in South Ostrobothnia. Central Ostrobothnia differs from its surrounding provinces based on PC4, PC5, PC6 and PC7 (fig. 9).

3.6. Principal component whole genome donor data

Principal components show variation between East and West. Figure 10 shows clear clustering in blood donors carrying variant compared to whole genome of donors who's not carrying the variant. PC1 and PC2 explained most of the variation (fig 10.) and were chosen for further investigation.



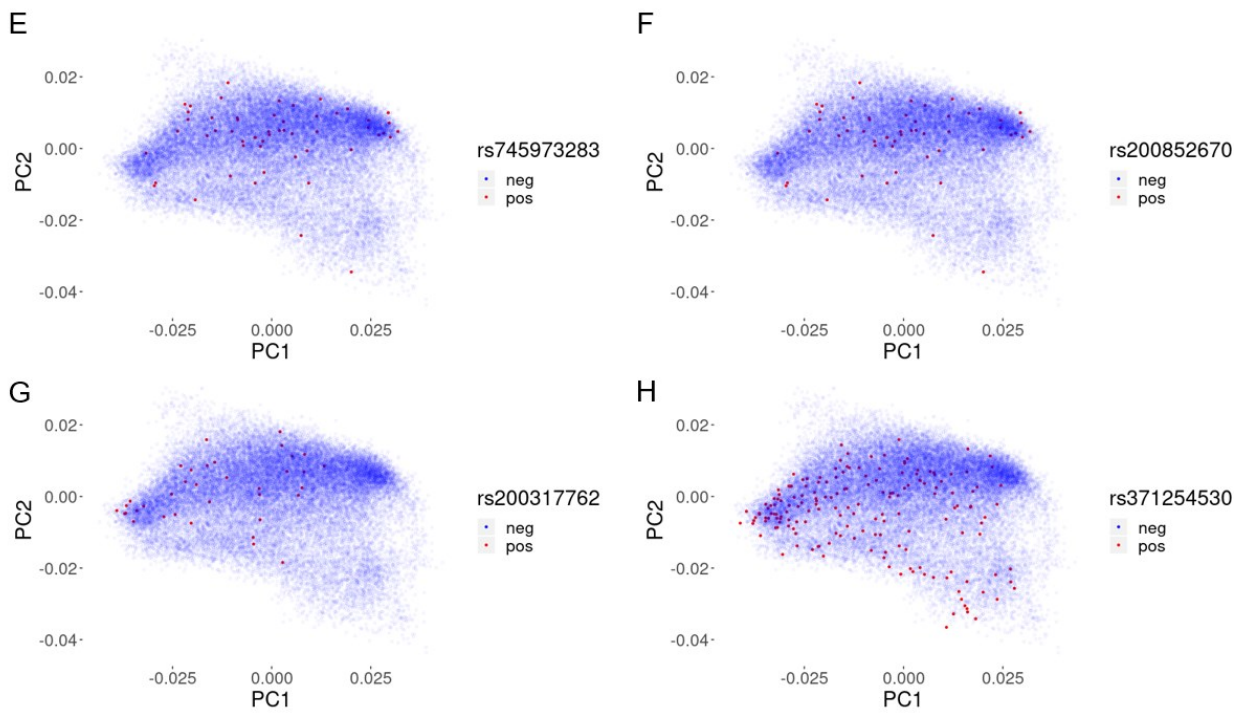


Figure 10. Principal components divided between variant negative and positive blood donors.

3.7. Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium was tested for all of the 51 variants individually (table 5). All of the variants are in Hardy-Weinberg equilibrium in the FRC Biobank data based on the q-values. The P-value of the two variants is lower than 0.05.

Table 5. Hardy-Weinberg exact test for the 51 variants

Hardy-Weinberg exact test	rs147660927	rs121909293	rs199935580	rs35937944	rs11591147	rs141266925	rs766868752	rs200852670	rs142351376	rs144109867
p-value	8.90E-01	8.98E-01	1.00E+00	4.2E-02	6.00E-01	8.57E-01	1.00E+00	1.00E+00	1.85E-01	1.00E+00
q-value	1.00E+00	1.00E+00	1.00E+00	5.53E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
homozygous donors (n)	32612	33917	35355	34177	32831	34349	35353	35344	34390	34083
heterozygous donors (n)	2730	1467	45	1205	2525	1043	47	56	999	1305
rare homozygous donors (n)	58	16	0	18	44	8	0	0	11	12
Hardy-Weinberg exact test	**rs200317762**	**rs148781286**	**rs201829738**	**rs149722682**	**rs201483470**	**rs41531245**	**rs74653330**	**rs147301839**	**rs147972626**	**rs201864074**
p-value	1.00E+00	7.29E-02	1.00E+00	1.00E+00	3.66E-02	7.19E-01	5.70E-01	6.12E-01	2.27E-01	3.79E-01
q-value	1.00E+00	9.20E-01	1.00E+00	1.00E+00	4.93E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
homozygous donors (n)	35228	35033	35186	35326	34897	34879	32355	34642	34607	34541
heterozygous donors (n)	172	364	214	74	498	519	2972	753	785	852
rare homozygous donors (n)	0	3	0	0	5	2	73	5	7	7
Hardy-Weinbrg exact test	**rs144651842**	**rs201162411**	**rs72553883**	**rs201955556**	**rs138213197**	**rs199598395**	**rs80338958**	**rs74006007**	**rs201208667**	**rs187429064**
p-value	8.79E-01	4.01E-01	8.22E-01	1.37E-01	4.30E-01	5.56E-01	1.00E+00	1.00E+00	6.54E-01	5.96E-02
q-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.69E-01
homozygous donors (n)	30265	34960	33700	34503	34897	34742	35358	34466	34955	34955
heterozygous donors (n)	4937	438	1681	895	503	654	42	928	443	3366
rare homozygous donors (n)	198	2	19	2	0	4	0	6	2	108
Hardy-Weinberg exact test	**rs184042322**	**rs150414818**	**rs371254530**	**rs199600574**	**rs200336521**	**rs77482050**	**rs145955907**	**rs201557719**	**rs780302457**	**rs74203920**
p-value	4.45E-01	1.00E+00	1.00E+00	1.00E+00	6.11E-01	6.07E-01	1.00E+00	2.12E-01	1.00E+00	9.36E-01
q-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
homozygous donors (n)	34664	34791	35379	34169	35036	31636	34051	35217	35349	32994
heterozygous donors (n)	731	607	21	1221	363	3653	1336	182	51	2365
rare homozygous donors (n)	5	2	0	10	1	111	13	1	0	41
Hardy-Weinberg exact test	**rs17879961**	**rs45620037**	**rs776981958**	**rs116483731**	**rs770636874**	**rs745973283**	**rs62621812**	**rs55960271**	**rs771807370**	**rs77273740**
p-value	3.84E-01	5.84E-01	1.00E+00	3.43E-01	1.00E+00	1.00E+00	1.00E+00	8.72E-01	1.00E+00	4.00E-01
q-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
homozygous donors (n)	33425	35049	35330	33379	35268	35363	32617	34226	35324	32200
heterozygous donors (n)	1942	350	70	1997	132	37	2727	1164	76	3117
rare homozygous donors (n)	33	1	0	24	0	0	56	10	0	83
Hardy-Weinberg exact test	**rs199680517**									
p-value	1.00E+00									
q-value	1.00E+00									
homozygous donors (n)	34123									
heterozygous donors (n)	1266									
rare homozygous donors (n)	11									

4. Discussion

In this project I investigated how well blood donor biobank which is considered healthy population will work for finding the identified rare variants enriched in Finland. The results showed that biobank is suitable for finding these disease associated variants enriched in Finland. All 51 variants were found among blood donor population, considered healthy, as heterozygous genotype. Most variants were also found to be homozygous, but some were not because in this size data set, very rare alleles only occur as heterozygous because the probability of homozygotes occurring is low. Results showed that some donors carry more than one variant and 20,4% of the donors don't carry any of the variants. Most of the variants showed higher frequency in Eastern provinces.

Allele frequency correlation tests showed strong correlation between blood donor Biobank and other two Finnish datasets (FinnGen and gnomAD-FIN). FinnGen dataset consists mainly of hospital biobank material from diseased individuals, for that reason a strong correlation between considered as healthy blood donor dataset and FinnGen dataset was not expected. There is a slight deviation correlation between FRC Biobank and FinnGen minor allele frequencies as FinnGen dataset contains genome data from FRC Biobank data and cannot be removed. The correlation between Biobank and non-Finnish gnomAD-NFSEE datasets was not as strong as the correlation between other two Finnish datasets. GnomAD-NFSEE reference dataset includes rest of the Europe (excludes Finnish-Swedish-Estonian populations) and minor allele frequencies were counted for variants enriched especially in the Finnish population. Excluding Swedish and Estonian samples from gnomAD-NFSEE dataset results in weaker correlation. This is because of the many chromosomes from sequencing studies of the Swedish population show Finnish origin, and Estonian population shares the same elements due the migration from Finland to Sweden in the 20th century (Kurki et al. 2022).

The findings show that 30 of the 51 variants differ between Eastern and Western provinces of Finland. Five of the 8 (table 4.) examined variants also showed statistically significant difference between a specific province compared to rest of the provinces in Finland. Principal component PC1 divide Finnish population genetically in two parts, Eastern and

Westerns Finland. The other principal components represent variation between the provinces. PC2 shows the difference between Ostrobothnia provinces in the West and the rest of the provinces. South Ostrobothnia differs from its surrounding provinces as well as Central Ostrobothnia in PC4, PC5, PC6 and PC7. Based on Salmela et al (2008) Finland's location between Scandinavian population from the West and Russian population from the East can still be seen in whole genome PCA results. Geopolitical location of Finland between two nations, Sweden and Russia, divided the population into two, which also reflects from results. Kerminen et al (2017) also states that the long-term influences may have affected to the main genetic division between East and West parts of Finland, which can also be seeing in blood donor PCA comparison.

The Finnish population history can explain the study results and variant distribution. Population genetic studies have shown that the Finns are part of the genetic continuity between mainland Europe and Uralic- speaking population from Siberia (Tambets et al. 2018). Genetic division between the East and West was nicely shown by Salmela et al 2008. by using genome data; Western Finns have Swedish component and Estonian and Eastern Finns have higher Siberian genetic component in their genome. Översti et al. (2017) described the Southwest costal area having connections to Scandinavia and inland areas having connections to the East. The differences between variants' distribution among FRC Biobank blood donor dataset in East and West provinces of Finland are in line with previous findings (Salmela et al. 2008; Tambets et al. 2018) and can be explained by population history. The 'Finnish Disease Heritage' illnesses have been described to be more common especially in the Eastern Finland (Kere 2001). This is supported by our findings that three out of four variants in blood donors, which differ most in terms of prevalence between eastern and western provinces (fig. 8) have higher frequency in the East than in the West. Out of all of the 51 variants 29 showed higher frequency in the East. Neuvonen et al. (2015) and Lappalainen et al. (2006) showed that the strong genetic border between West and East in Finland in both mitochondrial DNA and Y-chromosome data is exceptional in Europe which can explain variant distribution into two. The isolation of Finnish population caused by the geographical location, religious, and language boundaries have the enrichment of disease associated variants in Finland (Norio M et al. 1973), which explains why the 51

variants can be found among Finnish blood donors population compared to non-Finnish population (fig. 7C).

Further investigation of these results using only blood donors' whole genome PCA results, leaving out the postal code information, also showed clustering in PC1 and PC2. The division between East and West can be inferred from PCA results. These results also confirm the phenomenon in Finnish population history to still affect on today's Finnish population and are in line with the main population structure in Finland.

All of the investigated variants among blood donor population were in Hardy-Weinberg equilibrium indicating that natural selection does not have major impact on the variants. Thus variants have not been removed from the population which could explain why they have not been removed from the population by natural selection. Because of isolation, deleterious variants present relatively higher frequencies as a result of increased drift and reduced selective pressure (Casal et al. 2013). Natural selection has affected the tendency of disease associated risk alleles to be enriched in minor alleles in Mendelian diseases. Previously advantageous or neutral allele may later become a risk allele. (Klitz et al 1986). Minor allele is the second most common allele in a population (Kim et al. 2011). Kido et al. 2018 showed that the risk alleles of common diseases tend to be minor alleles because of minor alleles are more easily detected as risk allele in GWASs. Carrier with homozygous genotype is more likely to be affected by the variant because the carrier has inherited disease alleles from both parents. Variants are associated with genetic disorders which are also affected by environmental factors such as diet (Lobo 2008).

P-value adjustment did affect statistical results. Using the adjusted q-values for multiple testing reduced the false discovery rate and possibility of getting false significant results. As the examined variants cause genetic diseases, they may be affecting Finnish population also in the future, although their impact can be reduced by lifestyle factors, which contribute to the individual-level risk of diseases and individuals could stay healthy enough for blood donation.

Based on the project Biobank can focus the wanted variant sample collection in specific parts of Finland rather than collecting samples from all around the country. Biobank

provides material and data for research to investigate disease variants and pathogenesis. Collecting disease associated variants can provide important material for research. Further investigation is needed to find donors carrying more than one variant. Blood donor postal code information could cause slight error to the map visualization, as we cannot be sure that all of the donors still live in the same postal code area that they have informed during blood donation.

5. Conclusion

FRC Blood Service Biobank is a valuable resource for identifying blood donors with rare disease associated variants. Biobank can focus the sampling on specific provinces in Finland in the future, where a particular variant is present in higher prevalence. The number of variants varies considerably in blood donors, i.e. some carry none while some more than one. Variant distribution among blood donor population complies with the Finnish genetic population history.

6. Acknowledgment

I would like to show my appreciation to Jukka Partanen for offering me the Master of Science thesis project. I also want to thank my supervisors Jonna Clancy, Satu Koskela and Irma Saloheimo. Big thanks also go to Mikko Arvas, Jarmo Ritari and Jarkko Toivonen. Lastly, I'd like to mention Jussi Halonen and Amanda Sorvisto.

7. References

Blackburn A. N., Blondell L., Kos MZ., Blackburn NB., Peralta JM., Stevens PT., Lehman, DM., Blangero, J. & Göring HH. (2020). Genotype phasing in pedigrees using whole-genome sequence data. European Journal of Human Genetics, 28(6), 790-803. doi: 10.1038/s41431-020-0574-3.

Bloodservice. (2021). Privacy statement for blood service biobank register. [Referred 5.5.2022]. Available: https://www.bloodservice.fi/about-us/privacy-statement-for-blood-service-biobank-register

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300. doi:10.2307/2346101

Benjamini, Y. & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics, 29(4), 1165–1188. http://www.jstor.org/stable/2674075.

De La Chapelle, A. & Wright, FA. (1998). Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proceedings of the National Academy of Sciences, 95(21), 12416-12423.

Danecek, P., Auton, A. & Abecasis, G. (2011). The variant call format and VCFtools. Bioinformatics;27(15):2156-2158. doi:10.1093/bioinformatics/btr330.

FinnGen. (2022a). Biopankkien rooli. [Referred 6.5.2022]. Available: https://www.finngen.fi/fi/biopankkien_rooli.

FinnGen. (2022b). Genotypes. [Referred 10.6.2022]. Available: https://finngen.gitbook.io/documentation/methods/genotype-imputation.

FinnGen. (2022c). Genotype imputation. Last modified 2022. [Referred 10.6.2022]. Available: https://finngen.gitbook.io/documentation/methods/genotype-imputation/genotype-imputation.

FinnGen (n.d.). FinnGen project [Referred 10.5. 2022]. Available: https://www.finngen.fi/en/for_researchers.

FinnGen. (2022d). Sample QC and PCA. Last modified 2022. [Referred 10.6.2022]. Available: https://finngen.gitbook.io/documentation/methods/phewas/quality-checks.

FinnGen. (2022e). SISu reference panel. Last modified 2022. [Referred 10.6.2022]. Available: https://finngen.gitbook.io/documentation/methods/genotype-imputation/sisu-reference-panel.

GnomAD. Genom Aggregation Database. [Referred 3.4. 2022]. Available: https://gnomad.broadinstitute.org/.

Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Masson, G., Agnarsson, B. A., Benediktsdottir, K. R., ... & Stefansson, K. (2012). A study based on whole-genome

sequencing yields a rare variant at 8q24 associated with prostate cancer. Nature genetics, 44(12), 1326-1329.

Haggren, G., Halinen, P., Lavento, M., Raninen, S., & Wessman, A. (Eds.). (2015). Muinaisuutemme jäljet: Suomen esi-ja varhaishistoria kivikaudelta keskiajalle. Gaudeamus.

Haukka, J. (2022). Package 'mapsFinland'. [Referred 20.5.2022]. Available: https://cran.r-project.org/web/packages/mapsFinland/index.html

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. doi:10.1098/rsta.2015.0202

Kido, T., Sikora-Wohlfeld, W., Kawashima, M., Kikuchi, S., Kamatani, N., Patwardhan, A., ... & Butte, A. J. (2018). Are minor alleles more likely to be risk alleles?. BMC medical genomics, 11(1), 1-11. doi:10.1186/s12920-018-0322-5

Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. Restorative dentistry & endodontics, 42(2), 152-155. doi: 10.5395/rde.2017.42.2.152.

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., ... & Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC bioinformatics, 12(1), 1-16. doi:10.1186/1471-2105-12-23

Kenney, JF. & Keeping, ES. (1962). The Mode, Relation Between Mean Median Mode, and Relative Merits of Mean Median Mode. Mathematics of Statistics, Pt. 1, 3rd ed: 50-54, Princeton, NJ: Van Nostrand.

Kere, J. (2001). Human population genetics: lessons from Finland. Annual review of genomics and human genetics, 2(1), 103-128.

Kerminen, S., Havulinna, A. S., Hellenthal, G., Martin, A. R., Sarin, A. P., Perola, M., ... & Pirinen, M. (2017). Fine-scale genetic structure in Finland. G3: Genes, Genomes, Genetics, 7(10), 3459-3468. doi:10.1534/g3.117.300217.

Klitz, W., Thomson, G., & Baur, M. P. (1986). Contrasting evolutionary histories among tightly linked HLA loci. American journal of human genetics, 39(3), 340.

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K., ... & Nelis, M. (2022). FinnGen: Unique genetic insights from combining isolated population and national health register data. medRxiv. doi:10.1101/2022.03.03.22271360

Kääriäinen, H., Muilu, J., Perola, M. & Kristiansson, K. (2017). Genetics in an isolated population like Finland: a different basis for genomic medicine?. J Community Genet. 2017;8(4):319-326. doi:10.1007/s12687-017-0318-4

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., ... & Kattman, B. L. (2020). ClinVar: improvements to accessing data. Nucleic acids research, 48(D1), D835-D844. doi:10.1093/nar/gkz972.

Lappalainen, T., Koivumäki, S., Salmela, E., Huoponen, K., Sistonen, P., Savontaus, M. L. & Lahermo, P. (2006). Regional differences among the Finns: a Y-chromosomal perspective. Gene, 376(2), 207-215.

Lobo, I. (2008). Multifactorial inheritance and genetic disease. Nature Education 1(1):5

Mayo, O. (2008). A Century of Hardy–Weinberg Equilibrium. Twin Research and Human Genetics, 11(3), 249-256. doi:10.1375/twin.11.3.249.

Nevanlinna, HR. (1972). The Finnish population structure A genetic and genealogical study. Hereditas, 71: 195-235. doi:10.1111/j.1601-5223.1972.tb01021.x

Norio, R. (2003). The Finnish disease heritage III: the individual diseases. Human Genetics. 112:470–526

Norio, R. & Löytönen, M. (2002). The Finnish disease heritage. Fennia 180: 1–2, pp. 177–182. Helsinki. ISSN 0015-0010

Norio, R., Nevanlinna, HR. & Perheentupa, J. (1973). Hereditary diseases in Finland. Ann Clin Res 1973; 5: 109–141.

Palo, J. U., Ulmanen, I., Lukka, M., Ellonen, P. & Sajantila, A. (2009). Genetic markers and population history: Finland revisited. European Journal of Human Genetics, 17(10), 1336-1346. doi:10.1038/ejhg.2009.53

Peltonen L., Jalanko A. & Varilo T. (1999). Molecular Genetics the Finnish Disease Heritage, Human Molecular Genetics, Volume 8, Issue 10, 1999, Pages 1913–1923, doi:10.1093/hmg/8.10.1913

Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A. M., Rootsi, S., Honkola, T., ... & Metspalu, M. (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biology, 19(1), 1-20. doi:10.1186/s13059-018-1522-1.

Raivola, V., Snell, K., Helén, I. & Partanen, J. (2019). Attitudes of blood donors to their sample and data donation for biobanking. European Journal of Human Genetics, 27(11), 1659-1667. doi:10.1038/s41431-019-0434-1

Ringnér, M. (2008). What is principal component analysis?. Nature biotechnology, 26(3), 303-304. doi:10.1038/nbt0308-303

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: https://www.r-project.org/

RStudio Team. (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. Available: https://www.rstudio.com/

Sajantila, A., Salem, A. H., Savolainen, P., Bauer, K., Gierig, C. & Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proceedings of the National Academy of Sciences, 93(21), 12035-12039.

Salmela, E., Lappalainen, T., Fransson, I., Andersen, PM., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, ML., Schreiber, S., Kere, J. & Lahermo, P. (2008). Genome-wide

analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS One. 3(10):3519. doi:10.1371/journal.pone.0003519.

Samtools. (2022). Bcftools Manual Page. Last modified 21.2.2022. [Referred 10.4.22]. Available: https://samtools.github.io/bcftools/bcftools.html.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics, 78(4), 629-644. doi:10.1086/502802.

Suomen biopankit. (2022). Mikä on biopankki?. [Referred 5.5.2022]. Available: https://www.biopankki.fi/mika-on-biopankki/.

Uffelmann, E., Huang, QQ., Munung, NS., De Vries, J., Okada, Y., Martin, AR., ... & Posthuma, D. (2021). Genome-wide association studies. Nature Reviews Methods Primers, 1(1), 1-21. doi:10.1038/s43586-021-00056-9

Zhao JH. (2007). Gap: genetic analysis package. Journal of Statistical Software , 23(8):1-18. [Referred 3.6.2022]. Available: https://search.r-project.org/CRAN/refmans/HardyWeinberg/html/HWExact.html.

Översti, S., Onkamo, P., Stoljarova, M., Budowle, B., Sajantila, A., & Palo, J. U. (2017). Identification and analysis of mtDNA genomes attributed to Finns reveal long-stagnant demographic trends obscured in the total diversity. Scientific Reports, 7(1), 1-10.

Appendix 1.

````
---
title: "variant_filtering"
author: "Eevaleena Vaittinen"
date: "30 10 2022"
output: html_document
---

```{r}

library(tidyverse)
library(ggplot2)
library("xlsx")
```

#upload data
```{r}

dosage <- read.table("../dosage.tsv", header = TRUE, sep = "\t")
rsid <- read_excel("/variantid.xlsx")
```

#count hom, het, hom in total donors
```{r}

total <- dosage %>% select(-"donor") %>%
    gather(name,value) %>%  # reshape datset
    count(name, value) %>%
    spread(value,n) %>%
    rename(
      variant = "name",
      hom1 = '0|0',
      het1 = '0|1',
      het2 = '1|0',
      hom2 = '1|1')
total[is.na(total)] <- 0 #change NA to 0
total$het <- rowSums(total[3:4]) #sum het1 + het2 = het
total$total <- total$het + total$hom2 #sum donors carrying variant in total

```

#add variantids to the data frame
```{r}

total <- cbind(total, rsid)
total <- total[ ,-9] #remove unnecessary column

```

#filter all donors without variants
```{r}

x <- filter(dosage, variant1 == "0|0",  #example
                    variant2 == "0|0",
                    ...)
````
````

#plot heterozygous donors

```{r}

sum(total$het[1:51]) #count total number of heterozygous donors
ggplot(total, aes(x=variantid, y=het)) +
  geom_col(fill=)
  labs(title =,
       x="variants",
       y="donors") +
  theme_light() +
  theme(axis.title.y = element_text(hjust=0.5)) +
  theme(text = element_text(size = 15)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(axis.text.x = element_text(vjust = 0.5)) +
  geom_text()

```

#plot homozygous donors
```{r}

sum(total$hom2[1:51]) #count total number of homozygous donors

ggplot(totalMAF, aes(x=variant,y=hom2)) +
  geom_col(fill=) +
  labs(title =,
       x="variants",
       y="donors") +
  theme_light() +
  theme(axis.title.y = element_text(hjust=0.5)) +
  theme(text = element_text(size = 15)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(axis.text.x = element_text(vjust = 0.5)) +
  geom_text()

```

#write table
```{r}

write.table(total, file = "../total.tsv", sep = "\t",   #example
            quote = FALSE, row.names = FALSE)

```

Appendix 2.

````
---
title: "PCA_map"
output:
  word_document: default
  html_notebook: default
  pdf_document: default
  html_document:
    df_print: paged
---


```{r}
library(tidyverse)
library(ggplot2)
library(mapsFinland)

```

#upload data
```{r}

pca <- read.table("../pca.txt")
totalMAF <- read.table("../totalMAF.tsv", header = TRUE, sep = "\t")
dosage <- read.table("../dosage.tsv", header = TRUE, sep = "\t")
postcode <- read.table("../postcode.tsv", header = TRUE, sep = "\t")

```

#reshape dataframe (postcodes)
```{r}

postcode <- postcode[ ,-c(2:52)] #remove unnecessary columns
names(postcode)[names(postcode) == "donor"] <- "id" #rename column
postcode <- subset(postcode, PostalCode!="00000 TUNTEMATON POSTINUMERO" &
                             PostalCode!="00000 OBEKANT POSTNUMMER")
#remove donors without postal code

regionpca <- pca #new dataframe
names(regionpca) <- lapply(regionpca[1, ], as.character) #change headers
regionpca <- regionpca[-1,]
regionpca <- subset(regionpca, !is.na(IS_AFFY)) #remove donors without pc
results
regionpca <- regionpca[ ,c(1,4:23)] #remove unnecessary columns
regionpca <- regionpca %>% mutate_at(vars(starts_with("PC")), function(v)
as.numeric(as.character(v)))
summary(regionpca)
regionpca <- right_join(postcode, regionpca, by=c("id"="id")) #join
dataframes

```

#upload region postal code Excel data for every region
```{r}

region1 <- read_excel("../region1.xlsx") #postal codes from excel

```
````

#create new column for each region
```{r}

region$region <- "regionname"

```

#repeat for 19 regions

#create new dataframe with all of the reagions
```{r}

regions <- bind_rows(region1,region2...)
regionpca <- inner_join(regionpca, regions,
by=c("PostalCode"="PostalCode"))
colnames(regionpca)[23] <- "nimi" #rename column
regionpca <- regionpca %>% mutate_at(vars(starts_with("PC")), function(v)
as.numeric(as.character(v)))
summary(regionpca)

```

#count explained variance for every pca
```{r}

var(regionpca$PC1) / (var(regionpca$PC1)+var(regionpca$PC2)
+var(regionpca$PC3)+ var(regionpca$PC4)+var(regionpca$PC5)
+var(regionpca$PC6)+ var(regionpca$PC7)+ var(regionpca$PC8)
+var(regionpca$PC9)+ var(regionpca$PC10)+var(regionpca$PC11)
+var(regionpca$PC12)+var(regionpca$PC13)+ var(regionpca$PC14)+
var(regionpca$PC15)+ var(regionpca$PC16)+var(regionpca$PC17)
+var(regionpca$PC18)+var(regionpca$PC19)+ var(regionpca$PC20))

```

#repeat for every pca

#create data frame for explained variance
```{r}

PC <- c("PC1", "PC2","PC3"...)
var<- c(0.3,0.02, 0.001...)
pc_var <- data.frame(PC,var)
pc_var$PC <- factor(pc_var$PC, levels = pc_var$PC
                              %>% unique)

```

#plot pcs variance
```{r}

ggplot(pc_var, aes(x=PC, y=var, group = 1)) +
    geom_line(color = "red") +
    geom_point() +
   labs(title,
       y=,
       x=) +
  theme(axis.text.x = element_text(angle = 90))
```

```

#Change names
```{r}

colnames(regionpca)[23] <- "nimi"
```

#separte 10 first pca columns in one dataframe
```{r}
pc <- as.data.frame(regionpca[ ,c(1,3,4,5,6,7,8,9,10,11,12,23)])

```


#filter dataset and count median for each pca column
```{r}

md <- filter(pc, nimi == "regionname")
median(md$PC1)

```

#repeat for rest pca columns

#create dataframe for median results

```{r}

region <- c("regionname1","regionname2"..)
pcmedian <- c( 0.001, 0.002..)
pcmedian <- data.frame(region, pcmedian)

```

#repeat for rest PCs

#join dataframes to add map geometry
```{r}

map <- maakunta2019 #get the map
map <- x[ ,-c(1,2,4,5)] #remove columns
pcmedian <- right_join(pcmedian,map, by=c("regionname"="regionname"))

```

#plot pc median results to the map
```{r}

  ggplot(pcmedian, mapping = aes(geometry = geometry)) +
    geom_sf(aes(fill = pcmedian))+
    labs(fill = "pc median") +
    theme(legend.title = element_text(size = 15),legend.text =
element_text(size=15),
          legend.key.size = unit(10, "pt"),legend.key.width =
unit(10,"pt"),
          legend.key.height = unit(15,"pt"),
```
```

```
          legend.position = c(0.15,0.60)) +
    theme(axis.text.x = element_blank(),
          axis.text.y = element_blank(),
          axis.ticks = element_blank(),
          rect = element_blank()) +
    scale_fill_gradient(low="midnightblue", high="steelblue2")

```

#repeat for rest of the median pcas

#write table
```{r}

write.table(pca, "..PC/pca", sep = "\t",
            quote = FALSE, row.names = FALSE)

```

Appendix 3.

````
---
title: "MAF"
author: "Eevaleena Vaittinen"
date: "13 10 2022"
output: html_document
---
```{r}
library(tidyverse)
library(ggplot2)

```

#read allele frequency table
```{r}
MAF <- read.table(../"MAF.txt", header = TRUE)
```

#testing normal distribution for each dataset
```{r}
shapiro.test()
```

#correlation
```{r}
cor.test(data, method = "spearman", exact=FALSE)
Cor.test.MAF.FREK <- cor.test(data,method = "spearman")
```

#text string
```{r}
cor.text <- paste0("correlation: ", Cor.test.MAF.FREK$estimate %>%
signif(4), ' (',
                        Cor.test.MAF.FREK$conf.int[1] %>% signif(4), '-',
                        Cor.test.MAF.FREK$conf.int[2] %>% signif(4), ')\n',
                        ' p-value: ', Cor.test.MAF.FREK$p.value %>%
signif(3)

```

#scatter plot
```{r}
scatter_plot <- ggplot(data, aes(x=, y=))
scatter_plot + geom_point() + labs(x = "", y = "") +
  geom_smooth(method = "lm", col ="red") +
  theme(text = element_text(size = 15))

```
````

Appendix 4.

````
---
title: "HWE"
author: "Eevaleena Vaittinen"
date: "13 10 2022"
output:
  html_notebook: default
  pdf_document: default
---
```{r}

library(tidyverse)
library(HardyWeinberg)

```


#upload data
```{r}
data <- read.table("..data.tsv", header = TRUE, sep = "\t")
data <- data %>% mutate_at(vars(starts_with("chr")), function(v)
as.character(v))
sapply(data, class)

```

#Change values in dataframe
```{r}

data[data == "0|0"] <- "0"
data[data == "0|1"] <- "1"
data[data == "1|0"] <- "1"
data[data == "1|1"] <- "2"


```

#count AA, AB, BB genotypes
```{r}

for (i in names(data)[2:ncol(data)]) {

  x <- table(data[[i]])
  z <- as.vector(x)
  names(z) <- names(x)
  hwe <- print(z)
  hwe(z, data.type = "count", yates.correct=FALSE, miss.val=0)
}

```


#set genotype counts for variant
```{r}
hwe <- c(AA=32612,AB=2730,BB=58) #example

```
````

#count hwexact for each variant
```{r}
hwe <- HWExact(hwe)

```

#repeat

Appendix 5.

````
---
title: "pvalue_adjusment"
author: "Eevaleena Vaittinen"
date: "27 10 2022"
output: html_document
---

#set p-values
```{r}
pvalues <- c(0.001,0.002..) #example
```

#adjust p-values
```{r}

pvalues <- as.data.frame(p.adjust(pvalues, method = "BY"))
```
````

Appendix 6.

````
---
title: "chi2"
author: "Eevaleena Vaittinen"
date: "29 11 2022"
output: pdf_document
---

#create data frames for variants with carriers n > 30 compaired to donors
in east n=10585 and west n=24667
```{r}

snp1 <- data.frame(variant = c(978, 1795),
                   donors = c(9607,22872)) #example
rownames(snp1) <- c("east","west")
```

#run chi2 test
```{r}
chisq.test(snp1, correct = FALSE)
```

#create data frames for variants with carriers n < 30
```{r}
snp2 <- data.frame(variant = c(12,44),
                   donors = c(10573,24623)) #example
rownames(snp8) <- c("east","west")
```

#run fisher's exact test
```{r}
fisher.test(snp2)
```

#repeat
````

Appendix 7.

````
---
title: "donor"
author: "Eevaleena Vaittinen"
date: "10 10 2022"
output:
  word_document: default
  html_notebook: default
---
```{r}
library(tidyverse)
library(ggplot2)
```

#script to count donors in each 19 provinces

#example data set

 $ donor: Factor "id"
 $ chr  : Factor  "0|0","0|1","1|0", "1|1"

#upload data
```{r}

data <- read.table(file="data.tsv", header = TRUE, sep = "\t")
data2 <- "data.rdata"

```

#get the donor postal codes from donation data
```{r}

postalcode <- data2$donor
names(postalcode)[names(postalcode) == "id"] <- "donor"
location <- left_join(dosage,postalcode, by=c("donor"="donor")) #join based
on id
sum(is.na(location$PostalCode)) #count the number of donors without
postalcode in
location <- subset(location, !is.na(PostalCode)) #remove donors without or
unknown postalcode
location <- subset(location, PostalCode!="00000 TUNTEMATON POSTINUMERO" &
PostalCode!="00000 OBEKANT POSTNUMMER")

```

#get the postal codes for each region from Excel files
```{r}

region <- read_excel("../region.xlsx")

```



#Joining donors and postal codes
```{r}
region <- inner_join(data2, region, by=c("PostalCode"="PostalCode"))

```
````

#remove unnecessary columns from each dataframe
```{r}
region <- region[ ,-c(53:58)]

```

#count minor major homozygous and heterozygous genotypes
```{r}
maf <- region %>% select(-"donor") %>%
  gather(name,value) %>%  # reshape dataset
  count(name, value) %>%
  spread(value,n) %>%
  rename(
    variant = "name",
    hom1 = '0|0',
    het1 = '0|1',
    het2 = '1|0',
    hom2 = '1|1')

#change NA to 0
maf[is.na(maf)] <- 0
#sum het1 + het2 = het
maf$het <- rowSums(maf[3:4])

```

#repeat for all 19 regions

#count maf
```{r}
maf$hom <- maf$hom2*2
maf$maf <- maf$het + maf$hom
maf$maf <- maf$MAF/nro of donors x 2
```

#repeat for all regions

#Bind MAF columns in one data frame
```{r}
regionMAF <- bind_cols(maf,maf2..)
```

#remove unnecessary columns
```{r}
regionMAF <- regionMAF[ ,-c(2:7)]

```

#rotate and reshape dataframe in excel
```{r}
regionMAF <- t(regionMAF)
write.xlsx(regionMAF, file = "region.xlsx",
           sheetName = "Taul1", append = FALSE)

regionMAF <- read_excel("../regionMAF.xlsx")

regionMAF$variant <- factor(regionMAF$variant, levels = regionMAF$variant
                            %>% unique)
```

```

#Divide regions in to east and west
```{r}
regionMAFe <- regionMAF
regionMAFe$region <- "east"
regionMAFe <- regionMAFe[-c(1:10), ]

regionMAFw<- regionMAF
regionMAFw$region <- "west"
regionMAFw <- regionMAFw[-c(11:19), ]

MAFew <- bind_rows(regionMAFe,regionMAFw)
MAFew <- MAFew %>% mutate_at(vars(starts_with("chr")), function(v)
as.numeric(as.character(v)))
```

#Plot donors in total in each region

#creating new dataframe to plot donors in each region
```{r}
region <- c("region1", region2"...)
donors <- c(1, 2.. )
df <- data.frame(region, donors)
```

#plot number of donors in each region
```{r}
ggplot(df, aes(x=`region`, y=`donors`)) +
  geom_col(fill="orangered") +
  labs(title,
       y="donors",
       x="regions") +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(text = element_text(size = 30)) +
  theme(axis.text.x = element_text(vjust = 1))
```

#plot maf in east & west
```{r}
legend_title <- "variantid"
ggplot(MAFew, aes(x=region, y=position)) +
  geom_boxplot(aes(fill=MAFew$region), width=0.5)+
  ggtitle("Minor allele frequencies") +
  labs(title,
       x="Regions",
       y="") +
  theme(legend.title = element_text(size = 30),legend.text =
element_text(size=30),
        legend.key.size = unit(20, "pt"),legend.key.width = unit(10,"pt"),
        legend.key.height = unit(7,"pt")) +
  theme(axis.text.x = element_text(angle = 90)) +
  theme(axis.title.y = element_text(hjust=0.5)) +
  theme(text = element_text(size = 30)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(vjust = 0.5)) +
  theme(panel.background = element_rect(fill = "white")) +
  theme(panel.grid.major = element_line(size = 0.5, linetype = 'solid',

```
                                    colour = "grey")) +
  scale_fill_manual(legend_title,values=c("firebrick1","blue"))

```
#repeat for other variants

#plot maf in regions
```{r}
ggplot(MAFew, aes(x=variant, y=position)) +
  geom_col() +
  labs(title = "variantid",
       x="Regions",
       y="")+
  theme(axis.text.x = element_text(angle = 90)) +
  theme(axis.title.y = element_text(hjust=0.5)) +
  theme(text = element_text(size = 35)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(vjust = 0.5))

```
#repeat for other variants

#create data frame to count donors in east and west

```{r}
#summarize donor count in regions
nrow(region1)
nrow(region2)

east <- c("region1", "region2"..)
donors <- c(1,2...)
donorsE <- data.frame(east,donors)
sum(donorsE$donors)

west <- c("region3", "region4"..)
donors <- c(1,2...)
donorsW <- data.frame(west,donors)
sum(donorsW$donors)

#donors in east&west
WEST <- bind_rows(region1,region2..)
EAST <- bind_rows(region3,region4..)

```

#count genotypes
```{r}
MAFwest <- WEST %>% select(-"donor") %>%
  gather(name,value) %>%  # reshape datset
  count(name, value) %>%
  spread(value,n) %>%
  rename(
    variant = "name",
    hom1 = '0|0',
    het1 = '0|1',
    het2 = '1|0',
    hom2 = '1|1')
#change NA to 0
```

```
MAFwest[is.na(MAFwest)] <- 0
#sum het1 + het2 = het
MAFwest$het <- rowSums(MAFwest[3:4])
#count variants in total
MAFwest$total <- MAFwest$het + MAFwest$hom2

```

#repeat for other wind directions

```{r}
#count minor allele frequency
MAFeast$hom <- MAFeast$hom2*2
MAFeast$MAF <- MAFeast$het + MAFeast$hom
MAFeast$MAF <- MAFeast$MAF/21170

```

#write tables
```{r}
write.table(data, "../data.tsv", sep = "\t",
                        quote = FALSE, row.names = FALSE)
```