HelmholtzZentrum münchen
German Research Center for Environmental Health

Quality Control and Imputation of Genotype data from different Illumina arrays

Xenofon Giannoulis

A thesis presented for the degree of
Master of Science

Department of Computer Science and Medical Informatics
Faculty of Informatics
Thessaly University
01.04.2020

Thesis Approval

This thesis entitled **Quality Control and Imputation of genotype data from different Illumina arrays** by **Xenofon Giannoulis** is approved for the degree of **Master of Science**.

Examiners:

.....

.....

.....

.....

Supervisor:

.....

Date:

.....

Place:

.....

Abstract

During the last 15 years, the successful development of genome-wide associations studies (GWAS) led to an outburst of research to evaluate and identify genetic factors that influence human traits and diseases. However, the capability of GWAS to identify robust genetic associations depends on the quality of the data. Quality control (QC) procedure is typically used in GWAS because large scale experiments generate true results with a certain error rate which is minimized by identifying individuals and SNPs that introduce bias to the study. By using biologically relevant metrics as a proxy for quality, we are able to provide a clean data set for association testing between a phenotype and a trait of interest.

We have obtained peripheral whole blood sample to extract DNA for genotyping from 250 patients (104 males, 146 females) undergoing total joint replacement surgery from four different cohorts. One of the most computationally demanding challenges of this master thesis is merging cohorts from different array versions. Because of the complicated genetic assumptions of Illumina genotyping data, special handling of the initial files need to be made before proceeding to the QC filters. In order to generate a high-quality Illumina genotyping merged data-set, we extract only the common variants across the different array types. We present the application of rigorous QC with PLINK where 9 individuals and 10,354 SNPs did not pass the QC criteria. The cleaned data set proceed for imputation using the Haplotype Reference Consortium in the Michigan server, incorporating the Minimac 4 algorithm for imputation and Eagle v2.4 for phasing. The final output consist of 39,127,678 SNPs without any additional filter applied. The result of this thesis provides a detailed pipeline for the QC that can effectively detect outlier individuals and SPNs, focused on methods when dealing with different arrays types of Illumina Core Exome Chip.

Περίληψη

Κατά τη διάρκεια των τελευταίων 15 ετών, η επιτυχής ανάπτυξη μελετών σε επίπεδο γονιδιώματος οδήγησε σε έκρηξη έρευνας για την αξιολόγηση και τον εντοπισμό γενετικών παραγόντων που επηρεάζουν τα ανθρώπινα χαρακτηριστικά και ασθένειες. Ωστόσο, η ικανότητα της μεθόδου να αναγνωρίζει ισχυρούς γενετικούς συσχετισμούς εξαρτάται από την ποιότητα των δεδομένων. Η διαδικασία ελέγχου ποιότητας των δεδομένων χρησιμοποιείται συνήθως επειδή τα πειράματα μεγάλης κλίμακας δημιουργούν πραγματικά αποτελέσματα με ένα συγκεκριμένο ποσοστό σφάλματος που ελαχιστοποιείται με τον προσδιορισμό των παρογόντων που εισάγουν μεροληψία στη μελέτη. Χρησιμοποιώντας σχετικές στατιστικές μετρήσεις για τη διασφάλιση της ποιότητας των δεδομένων, είμαστε σε θέση να παρέχουμε ένα καθαρό σύνολο δεδομένων για συσχέτιση δοκιμών μεταξύ ενός φαινοτύπου και ενός χαρακτηριστικού γονότυπου.

Λάβαμε δείγμα περιφερικού ολικού αίματος για να εξαγάγουμε ΔΝΑ από 250 ασθενείς (104 άνδρες, 146 γυναίκες) που υποβλήθηκαν σε ολική χειρουργική επέμβαση αντικατάστασης αρθρώσεων από τέσσερις διαφορετικές ομάδες. Μία από τις πιο απαιτητικά υπολογιστικές προκλήσεις αυτής της διατριβής είναι η συγχώνευση ομάδων από διαφορετικές εκδόσεις μικροσυστοιχιών. Λόγω των περίπλοκων γενετικών υποθέσεων των δεδομένων γονότυπου, πρέπει να γίνει ειδικός χειρισμός των αρχικών αρχείων πριν προχωρήσουμε στα εφαρμογή των φίλτρων ελέγχου ποιότητας. Προκειμένου να δημιουργήσουμε ένα συγχωνευμένο σύνολο δεδομένων υψηλής ποιότητας, εξάγουμε μόνο τα κοινά χαρακτηριστικά μεταξύ των διαφόρων τύπων συστοιχιών. Παρουσιάζουμε την εφαρμογή ποιοτικού ελέγχου δεδομένων όπου 9 άτομα και 10.354 νουκλεοτιδικοί πολυμορφισμοί δεν πέρασαν τα κριτήρια. Με το εναπομείναντα σύνολο δεδομένων εφαρμόζουμε μεθόδους υποκατάστασης ελλιπών δεδομένων ενσωματώνοντας διάφορους αλγόριθμους. Το τελικό σύνολο δεδομένων αποτελείται από 39.127.678 νουκλεοτιδικούς πολυμορφισμούς χωρίς να έχουμε κάποιο επιπρόσθετο φίλτρο. Το αποτέλεσμα αυτής της διατριβής παρέχει έναν λεπτομερή οδηγό για την εφαρμογή ποιοτικού ελέγχου δεδομένων κατά τη διάρκεια κοινών μελετώς συσχέτισης ολόκληρου γονιδιώματος που μπορεί να ανιχνεύσει αποτελεσματικά δείγματα και πολυμορφισμούς, λαμβάνοντας υπ'όψην διαφορετικές εκδόσεις συστοιχιών.

This thesis is dedicated to my parents
For their unwavering love, support and encouragement

Statement of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work and that all the assistance received in preparing this thesis and sources have been acknowledged. This thesis has not been submitted for any degree or other purposes.

Acknowledgments

First of all I would like to extend my heartfelt gratitude to my supervisor, Professor Artemis Hatzigeorgiou for her guidance and mentorship throughout my studies. I would particularly like to express my deepest appreciation to Professor Eleftheria Zeggini for giving me the opportunity to join the Institute of Translational Genomics in Helmholtz Centre for Environmental Health in Munich and implement the current master thesis. I have received generous support from Dr. Konstantinos Hatzikotoulas and I am thankful for his excellent supervision, and for the valuable comments and inputs. I am also grateful to my colleagues at Helmholtz, particular to Ms. Lorraine Southam, Dr. Rayner William, Dr. Arthur Gilly, and the Dr.-to-be Peter Kreitmaier, whose help was truly invaluable to complete this thesis. Furthermore, I acknowledge the efforts of the reviewers who have taken the time to read this thesis.

List of Abbreviations

APEX Array primer extension Assay
eQTL Expression of quantitative trait locus
DNA Deoxyribonucleic acid
GRCh37, hg19 Genome reference consortium, build 37
GWAS Gene wide association study
HRC Haplotype reference consortium
IBD Identify by descent
IBS Identify by state
HRC Haplotype reference consortium
HWE Hardy-Weinberg equilibrium
LD Linkage disequilibrium
MAC Minor allele count
MAF Minor allele frequency
MDS Multidimensional scaling
MIS Michigan imputation server
MT Mitochondrial
OA Osteoarthritis
PAR Pseudo autosomal region
PCA Principal Component analysis
SNP Single-nucleotide polymorphism
QC Quality control
1KG 1000 genome project

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 GWAS

GWAS have identified thousands of loci (Welter et al., 2014) and candidate regions playing an important role in complex traits and diseases (König et al., 2013). During the past years, GWAS analyses used to detect associations between genetic variants and genetic traits in population samples (Visscher et al., 2017), leading to a better understanding of the disease risk prediction and enabling preventative, population and personalized medicine. In addition, quantitative traits such as body mass index (Locke et al., 2015), cholesterol (Global Lipids et al., 2013) and human height (Lango Allen et al., 2010), helped researches reveal pathways in disease aetiology end evidence for underlying molecular mechanisms.

These efforts have remarkable increased genetic basis knowledge in a wide spectrum of disorders and diseases such as osteoarthritis (Tachmazidou et al., 2019), psychiatric (Horwitz et al., 2018), breast cancer (Fachal et al., 2020), heart disease (Shah et al., 2020) and others, making Homo sapiens the most phenotypically studied organism (Stranger et al., 2011). Since the first human genome sequence in 2003, almost 3700 (Patron et al., 2019) GWAS related publications were published (figure 1), across multiple data-sets and cases/controls from fluctuating ancestral diversity that looks at the genetic contribution of SNPs, identifying more than 10,000 loci (L. Duncan, A. Brown 2018) of interest.

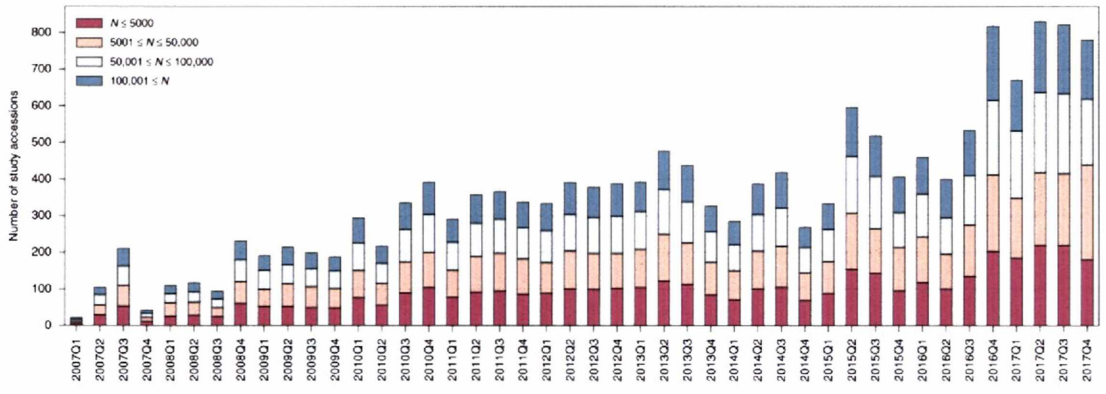

Figure 1: A scientometric review of GWAS studies development within the span of 10.years (2007-2017).Mills, M.C., Rahal, C. A scientometric review of genome-wide association studies. Commun Biol 2,9 (2019) doi:10.1038/s42003-018-0261-x.

The major challenges GWAS studies are facing are two. Firstly, because of the study sizes GWAS require, the rare variants or complex traits do not take under consideration the context of environmental risk factors which if included, could lead to a better explanation of the missing heritability (Frazer et al., 2009). Secondly, GWAS do not necessarily identify the causal variants (Altshuler et al., 2008), but usually define a pinpoint loci that possible causal variants lie. Moreover, at the moment GWAS signals have a limitation about predictive clinical treatment and the subject of prevention and treatment is currently under debate (Tam et al., 2019). Nevertheless, GWAS have contributed between others in the discovery of novel biological mechanisms, easy-to-share and publicly available data and remarkable insight into the ethic variation of complex traits.

1.2 Single Nucleotide Polymorphism (SNPs)

SNPs are the most common type of genetic variation in the human genome, defined as loci with alleles that differ at a single base (P.Daiger et al., 2013). It is possible to have two or more different nucleotides at a specific position on a chromosome. Although a set of SNP may not directly cause a disease, it is established that some specific markers are associated with certain disorders or a particular trait locus. These associations allow us to identify an estimation of an individual's genetic predisposition to develop a disease or not. Moreover, even if the most variation is found across all human populations, some variants appear to have a higher population or ancestry specific density. Studies in isolated populations reveal high impact insides about an under investigation phenotype.

There are different types of SNPs, since a single nucleotide may be changed (substitution), or being defined as indel, which is a term for insertion or deletion of bases in the genome of an organism or a polynucleotide sequence. They may also fall within non-coding regions of genes, where progress in genomic, transcriptomic and epigenomic profiling improve the prediction of the variant outcomes (S. Gloss and E. Dingere et al., 2018). Chip-based DNA genotyping allows the genotyping assays of a greater than 1 million SNPs simultaneously on one individual with a cost-effective procedure (M. Cotton et al., 2018). There has been a significant rise of the genomic analysis worldwide, leading to identifying roughly more than 100 million SNPs in populations around the world, available in the public dbSNP catalogue (A. Auton et al., 2015) (https://www.ncbi.nlm.nih.gov/snp/).

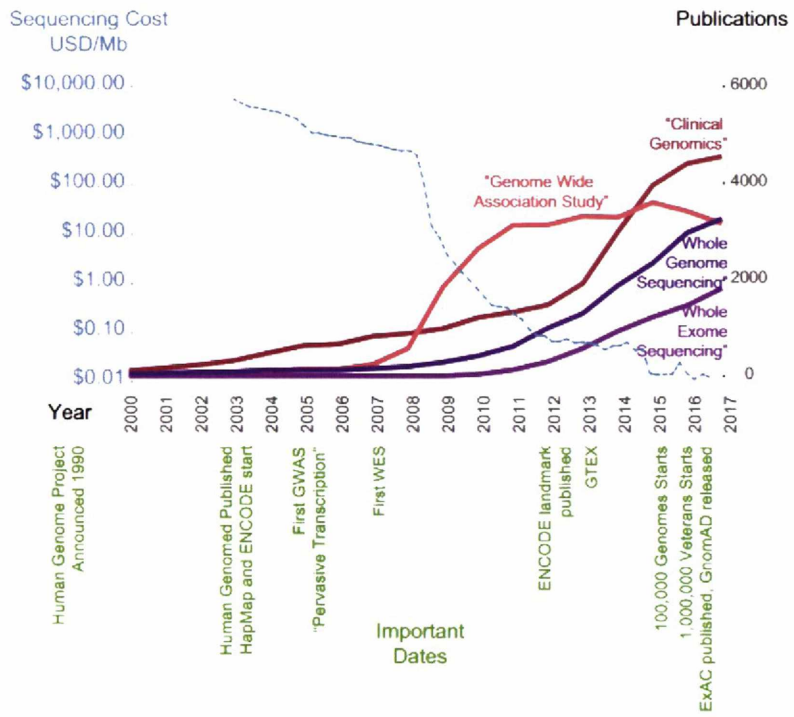

Figure 2: .
DNA sequencing costs over time compared to the number of publications (Gloss, B.S., Dinger, M.E. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med 50, 97 (2018). https://doi.org/10.1038/s12276-018-0087-0).

1.3 SNP genotyping arrays

SNP arrays are used to genotype human DNA at thousands of SNPs across the genome at a time or ideally simultaneously. Practically, an ''array" is a DNA sequence fixed on microscope slides or chips by forming chemical bonds (A. Missoum., 2018). From a certain sample, we can identify the presence of fluorescence nucleic acids by using a specialized scanner and software. There are different types of DNA microarrays that can be used to determine gene expression levels and SNP genotyping. Each SNP is independently analyzed to identify genotypes. Genotype calling for small-scale data efforts can be manually handled by the inspection of the allele-probe intensities.

For large-scale studies, like GWAS, it is widely used automated genotype calling algorithms, for example, Illumina Genome Studio software application (`https://www.illumina.com/techniques/microarrays/array-data-analysis-ex`

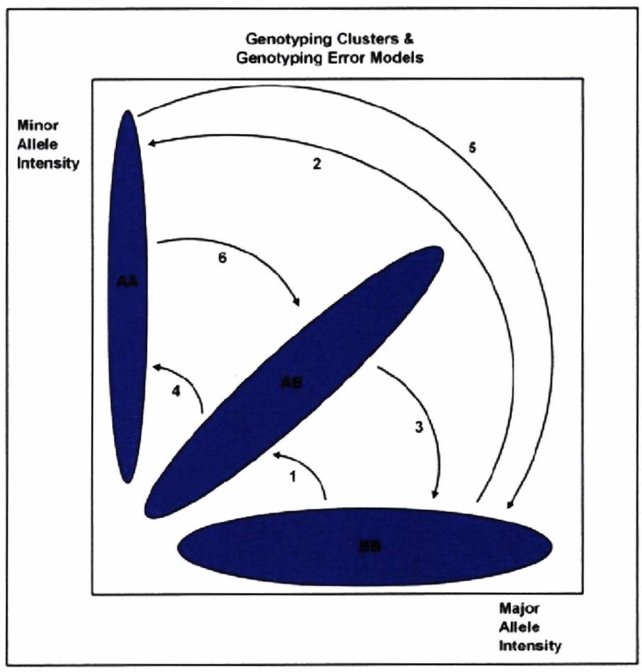

Figure 3: Generating intensity plots of the three genotype clusters. Each arrow represents any possible genotyping error combination. For example, in mode 6 minor homozygotes (AA) can be miscalled as heterozygotes (AB). Fardo DW, Ionita-Laza I, Lange C (2009) On Quality Control Measures in Genome-Wide Association Studies: A Test to Assess the Genotyping Quality of Individual Probands in Family-Based Association Studies and an Application to the HapMap Data. PLOS Genetics 5(7): e1000572. https://doi.org/10.1371/journal.pgen.1000572

perimental-design/genomestudio.html). There is a pipeline available with an extensive list of QC screening criteria (Grove et al., 2013) and Genome Studio procedures tailed to the Human Exome chip (Guo et al., 2014). The level of fluorescent intensity of each probe in genotyping raw fluorescence data represents the signal strength of each allele (Zhao et al., 2017). We have used the Gencall application which incorporates the GenTrain clustering algorithm and a calling algorithm analyzing raw intensity data for a given individual and a given marker locus. They estimate the probability that their genotype is aa, Aa, AA. In addition, the calling operation is performed by using a Bayesian model and a GenCall score which is the GenTrain Score and a data-to-model fit score, where a score below 0.2 is considered as failed call rate and a score above 0.7 is reposted as well-behaving genotype.

There are various companies providing array genotyping, including Illumina (https://www.illumina.com/techniques/microarrays.html), Affymetrix microarrays

Exome Type	Exome Version	Sample Number	Build
Human Core Exome	12v1-1_A	12	37
Infinium Core Exome	24v1-1_A	122	37
Infinium Core Exome	24v1-3_A1	116	37

Table 1: Exome and build version of the under investigation datasets, providing information about the sample size, Human exome Beadchip type of each array, and the strand build the samples are based.

(Affymetrix Inc., Santa Clara, CA), Array Primer Extension Assay (APEX), GenoSNP and CRLMM ensuring genotype call accuracy with the independent HapMap (`https://www.genome.gov/10001688/international-hapmap-project`) project as the gold standard. In our study, we have used the Illumina Infinium genotyping array, and specifically the Human Core Exome, using the GenCall array calling. The number of initial samples are 250, distributed in three different array types (table 1).

1.4 Elements of Quality Control

Quality control (QC) is a critical application during any genotype analysis. It is a method to identify DNA samples and markers that introduce bias in each genetic study. It takes place before any association test since even a small percentage of genotyping error lead to spurious results. This is the stage where we ideally identify sample errors due to DNA contamination, low quality or quantity of DNA, inbreeding, chemical or machinery failure and human mistakes during the DNA extraction. Statistical methods have been developed to ensure genotyping quality (Plagnol et al., 2007). QC covers two major areas, individual-based quality measures and variant-based quality measures, both to ensure that poor quality calling rates of individuals and SNPs are removed from further analysis. Before applying the QC criteria for each category, defining the sample genotyping efficiency is necessary by removing all samples and SNPs with more than 10% missingness. The thresholds within the individual and SNP QC are adjustable based on the sample size, the quality of the DNA samples used, and the variability in human and equipment error during the genotyping procedure.

1.4.1 Individual-based QC

Individual Call/Missingness Rate, –missing

Missingness which is the complement of call rate is the proportion of missing genotypes after algorithm calling per individual. It is an indicator of SNP quality from the original genotyping array and usually plotted along with heterozygosity rate because they affect each other and define the quality of the genotyped sample. A recommended threshold for removing individuals is the one with less than 98% call rate or more than 2% missingness.

Heterozygosity Minor Allele Frequency (MAF) – het, –maf

Heterozygosity is fundamental to the study of genetic variation in populations. It is the degree of having two different alleles at a locus. Excess heterozygosity indicates sample contamination (Jun et al., 2012) where sample inbreeding is the most common causal reason for low heterozygosity rates (Anderson et al., 2010). Inbreeding refers to mating between related individuals. The threshold used to remove individuals was 3 standard deviations from the heterozygosity mean rate of all samples. In addition, we combine heterozygosity with the MAF, which refers to the frequency at which the second most common allele occurs in a given population. In other words, if there are 3 genotypes with frequencies of 0.50, 0.36 and 0.01, the MAF will be reported as 0.36. In this step, MAF provides information that distinguish common polymorphisms (MAF1) from rare variants (MAF<1). It is common that genotyping individual call rate and heterozygosity call rate to be plotted together and cut-offs are selected so as to identify outlier individuals based on both statistical indicators.

Sex discrepancy –sex-check

Identifies that self-reported sex match the genotype sex, based on heterozygosity rate because males have a single X-chromosome and therefore can be estimated to be homozygous across all X-chromosome SNPs other than those in the pseudo-autosomal region. Therefore, X chromosome heterozygosity estimate for males is 1 and for females 0. Sample relatedness –genome Relatedness identifies possible cryptic relationships between individuals thought to be unrelated, sample identity errors in family data, improperly duplicated samples and contaminated samples that have escaped other screenings (O. Igo et al., 2016). Mendel's law (1866) for diploid organisms say that at any given locus, each individual has two genes, one

Pihat value	Interpretation
1	Duplicate sample or monozygotic twins
0.5	1rst degree relative
0.25	2nd degree relative
0.125	3rd degree relative

Table 2: Pihat values and interpretation

maternal and one paternal. Thus, individuals carry at a locus, pieces of DNA that are copied through repeated segregations from their ancestors.

Identical by Descent (IBD)

Identity by Descent (IBD) measures recent shared ancestry between two individuals or duplicate entries. Relatives who share a common ancestor may both carry copies of the sample ancestral piece of DNA. IBD has three coefficients (Z0, Z1, Z2) which are automatically produced when running –genome command on PLINK and estimate the probability two individuals sharing 0, 1 or 2 alleles. Using this probability model and method we estimate the Pr (Z) for any given marker. Some pedigree errors can be corrected by consulting original records, while others are corrected based on the inferred genetic relationship (Laurie et al., 2010). IBD is calculated and denoted in Plink as pihat, which is the threshold measures all pairs of individuals based on a SNP set. We used one of the most common thresholds in GWAS, pihat ¡ 0.2 to reassure that there is no ancestry relation between the individuals. Each pihat value corresponds to a degree of relatedness (Table 2). To retain the maximum number of test samples and thus statistical power, methods such as mixed regression model (Aulchenko YS et al., 2007) have been developed which account for that confounder in GWAS.

Identical by state (IBS) —mds

Identify by state (IBS) is used to calculate for each pair of individuals the average proportion of alleles shared in common at genotyped SNPs (including only the autosomal regions). Because this method works better when only independent SNPs are included, complex regions of extended linkage disequilibrium (LD) are removed from the merged dataset (Purcell S, et al., 2007). The remaining regions are pruned so that no pair of SNPs within a given region is correlated (r2¿0.2). When comparing Pedigrees and populations (IBD vs IBS), IBS is a reflection of IBD in both comparison cases. The main difference between pedigrees and popu-

lations is that a pedigree shares a very strong prior on IBD. Apart from the IBD coefficients, pihat is the main measure that helps us make a decision about which samples are duplicated/related or not.

Population structure and stratification –cluster

Multidimensional scaling (MDS) is a popular statistical method commonly used to visualize multidimensional datasets. MDS optimizes the following objective function:

$$\mathrm{F}_{MDS} = \sum_{i \neq j} [d(i,j) - d'(i,j)]^2,$$

where d(i,j) the distance between i and j in the original multidimensional space and d'(i,j) the distance between i and j in the projected 2D space. Population stratification in genetic association studies must be considered in the QC design to avoid spurious association or reduced power. The most common methods used in GWAS are genomic control (GC), EIGENSTRAT, principal component-based logistic regression (PCA-L),LAPSTRUCT, ROADTRIPS, and EMMAX. We used Principal Component Analysis (PCA) with PLINK 1.9 and the MDS based on raw Hamming distances. In addition, we merge our data with the 1000 Genome Project (1KG, https://www.internationalgenome.org/) to compare the population structure of the under investigation dataset with the representative collection of the 1KG reference panel. We kept only the overlapping autosomal variants across the merged file.

Principal Component Analysis (PCA)

When analysing genetic data, it is essential to determine individuals coming from an underlying structured population. In order to detect and quantify this structure and uncover the demographic history of under investigation population, we use the multivariate statistical method PCA. This method is widely used as a computationally efficient approach to deal with sample heterogeneity (Zhu et al., 2002; Novembre and Stephens, 2008). PCA is a non-parametric technique which uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The first resulting new variables (principal components) account for as much of the available information as possible. The first principal component has the larger possible variant across the complete dataset. The main aim of the method is the reduction, interpretation, and graphical representation

of the data and is commonly used among others in regression analysis to remove multi-collinearity between the covariates of the model and in cluster analysis, to avoid giving too much weight to correlated variables in the definition of the clusters.

Practically, the method is applied during QC of genetic data at a population level in combination with the under investigation individuals' cluster set, to identify ethnicity outliers. In order to have a reference panel to compare the under investigation dataset with other ethnicities, we have used the 1000 Genome Project (1KG) (https://www.internationalgenome.org/data/) which provides a public resource of genetic variation to support the next generation of medical association studies. This reference panel contains both common and rare variants across the genome and the geographical origin of 14 populations (phase 1; 1092 individuals). The 1KG project has led to a huge contribution to the cost reduction of the human genome sequencing and the discovery of new SNPs.

1.4.2 SNP-based QC

The main aim of SNP-Based QC is to identify and remove poor quality SNPs that can lead to spurious results. Here, we prior exclude all individuals that did not pass the criteria of individual-based QC and split the markers into the following three categories for a thorough investigation of marker call rate and deviation of the Hardy-Weinberg Equilibrium principle:

- SNPs that belong to the autosomal region, chromosome 1-22 (–autosome)
- SNPs that belong to males and females on the chromosome 25 (–chr 25)
- SNPs that belong to females in the chromosome 23 (–chr 23 –filter-females)

Marker call/Missingness rate –missing

This QC step checks the SNP genotyping efficiency and quality by calculating the proportion of individuals for whom information on a specific SNP is missing. The higher the missingness the lower the genotyping quality of the SNP. In this QC protocol, we removed all SNPs with call rate<2%.

Hardy-Weinberg Equilibrium (HWE) –hardy

We Identify SNPs demonstrating significant deviation from the Hardy-Weinberg principle, aiming to keep only quality markers and avoid spurious SNP call rates downstream. The assumptions of HWE are the following:

- There is an expected relationship between allele and genotype frequencies
- Allele frequencies and genotypes remain constant over non-overlapping generations
- Deviations from HWE are used as a proxy for possible genotyping errors
- Refers to diploid organisms and an infinite population size
- A population that satisfies the HWE principles, the two alleles on two chromosomes carried by an individual are independent of each other and a SNP with allele frequencies p and q, the three genotypes should have frequencies $p^2, 2pq, q^2$respectively.

Apart from any violations of the above assumptions, the most common reasons for Hardy-Weinberg extreme deviation are due to genotyping error, subdivided population, heterogeneity in genetics background and purely statistical chance.

MAF –maf

Markers with low minor allele count are often removed as the genotype calling is very difficult due to the small size of the heterozygote and rare-homozygote clusters. In this step, we calculate the minor allele frequencies for all variants passed the per-individual QC. No MAF filter was applied in our work as this step strongly relies on the particular needs of each study design.

1.5 The role of Genotype Imputation

Since the main benefits of the Human Core Exome is the low-cost array genotyping, to take the maximum advantage of its application, imputation of the data to a reference population is required, to allow rare coding variation to be assayed to a large number of samples without the costs of whole-genome sequencing these variants (Wagner MJ. Et al., 2013). This method is estimating genotype probabilities or genotypes at SNPs that have not been directly genotyped and practically increases the chances of identifying a causal variant in a GWAS. It is also a helpful tool for meta-analysis by facilitating the combination of results across studies (Sayantan Das., 2017).

1.5.1 The Haplotype Reference Consortium (HRC)

The haplotype is the combination of alleles on the same chromosome position. HRC (http://www.haplotype-reference-consortium.org/) is a predominantly European ancestry reference panel consisted of 64,976 human haplotypes and 39,235,157 SNPs with a minor allele count (MAC) greater or equal to 5 (McCarthy et al., 2016).

1.5.2 Pipeline of QC for imputation in the Michigan Imputation Server (MIS)

Before the actual imputation and the recovery of missing variants, Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html!pages/home) performs several pre-imputation QC steps to determine the number of valid variants:

- Chunk exclusion (variants<3 — overlap<50% — sample call rate<50%)
- Variant exclusion (invalid alleles occur, finding duplicates, indels, monomorphic sites, allele mismatch between the reference panel and study, SNP call rate<90%)
- Sample exclusion (if in for autosomal region samples with call rate<50% are found, complete chunks are removed, perform lift Over-step in case the input data and the reference panel does not match).

1.5.3 Phasing and Eagle software

Michigan Imputation Server (MIS) provides the option of phasing with Eagle software which works with the determination of the haplotype phase within a genotyped cohort or by using a phased reference panel. (Browning SR, 2011). During the last years, haplotype phasing methods have been developed to handle whole-chromosome data efficiently in the field of large-scale sequencing. It has been proved that by using haplotype phasing, our understanding in disease and variation (Tewhey R et al., 2011) is increased, by imputing missing genetic variation (Tewhey R, Browning SR., 2009). At the same time, when dealing with sequence and microarray data, phasing detects possible genotype errors during genotype calling (Kang H et al., 2004) (Scheet P, Stephens M, 2008), inferring human geographically dispersed populations (Tishkoff SA, et al., 1996), and assists to detect recurrent mutation (Kong A, et al., 2008).

MIS provides the option to use a reference-based haplotype phasing with Eagle (https://data.broadinstitute.org/alkesgroup/Eagle) and uses a Hidden Markov Model (HMM) algorithm to improve the accuracy and speed. The main advantage of using Eagle is the efficient haplotype matching by using the Burrows-Wheeler transformation (PBWT) (Durbin R., 2014), and Hidden Markov Model (HMM) algorithms which rapidly search for the most relevant paths.

1.6 Osteoarthritis

Osteoarthritis (OA), commonly known as wear-and-tear arthritis, is one of the most disabling diseases, affecting more than 240 million people worldwide (OARSI). It is mainly caused by damage or breakdown of joint cartilage, causing significant pain and moving disability. Common comorbidities of OA are between other cardiovascular diseases (Jeong et al. 2013), depression and mental disorders (Reeuwijk et al., 2010). The most common person-level risk factors of OA are obesity, occupation, gender, family history and Joint abnormities. Knee osteoarthritis long existed at low frequencies, but since the mid-20th century, the disease has doubled in prevalence (Wallace IJ et al., 2017). This fact highlights the need to further study preventable risk factors and possible associated causing targets that have become prevalent within the last century. Another aspect is that to date, there is no cure for OA and is considered as a heterogeneous disease (from a phenotype perspective) with very complex pathology.

Chapter 2

Materials and Methods

2.1 DNA extraction and storage condition

2.1.1 Data reporting

No statistical methods applied to predetermine the sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Furthermore, all patients provided written consent prior to participation in the study.

2.1.2 Sample collection

We have obtained peripheral whole blood sample to extract DNA for genotyping from 250 patients (104 males, 146 females) undergoing total joint replacement surgery in 4 cohorts, where after careful consideration of the manifest and the study design requirement to remove any patient undergoing hip replacement, we kept 220 for QC and imputation. These samples are distributed as follows: 12 knee osteoarthritis patients (cohort 1; 10 males, 2 females); 19 knee osteoarthritis patients (cohort 2; 6 males, 13 females) 73 osteoarthritis patients (cohort 3; 29 males, 44 females), and 116 osteoarthritis patients (cohort 4; 49 males, 67 females). In the graphical representation of the QC results, cohort 1 is mentioned as FunGen1, cohort 2 as FunGen2, cohort 3 as FunGenQQ and cohort 4 as FunGenQQ2.

2.1.3 DNA Extraction

DNA extraction was carried out using Qiagen Allprep Minikit (`https://www.qiagen.com/de/products/diagnostics-and-clinical-research/sample-processing/allprep-dnarna-mini-kit/`) following the instructions of the manufacturer, with no significant variations between the cohorts. Ethanol precipitation is used for concentrating and desalting nucleic acid, while DNA samples were frozen at -80o C. DNA extraction and 16 genotyping was carried out at the Welcome Trust Sanger Institute, Hinxton, UK (https://www.sanger.ac.uk/).

2.2 Software and Tools

2.2.1 Tools used

Due to the high computational power each step requires, LINUX environment is highly recommended running analyses and combine additional tools and plugins. A fast and user-friendly genetic tool-set is PLINK (`http://pngu.mgh.harvard.edu/purcell/plink/index.shtml`), which was designed by Shaun Purcell. The v1.90b6.13 64-bit (30 Nov 2019) version of PLINK was used for the analyses. The software performs a wide range of large scale analyses, implements QC procedures and statistical testing with customized thresholds. Moreover, analyses like population stratification, PCA and Linkage Disequilibrium (LD) calculations are flexible and precisely done. For the visualization, statistical summaries and analyses that cannot be run by PLINK, R statistical programming language (https://cran.r-project.org/) is an excellent tool for interpreting results and graphically present the findings of each study in an elegant way. The version of R used is 3.6.1 (2019-07-05). The imputation method was applied by using the Michigan Imputation Server 1.2.4 to obtain the maximum number of haplotypes and imputed variants. The imputation incorporates the Minimac 4 algorithm (1.2.4) and Eagle v2.4 for phasing. As reference panel, HRC r1.1 2016, GRCh37—hg19 was used. The software used is open-source and freely available on the provided links.

2.2.2 PLINK Annotation

Plink offers an effective way of data structure and efficiency. The .ped file contains both phenotype and genotype information and the .map file contains the position information of the genetic variants. The binary format (b-file) has the phenotypic and genetic information in 3 files (.fam, .bed, .bim). The .fam file contains the individuals' phenotype information, the .bed file the genotype data on a binary format and the .bim file contains the alleles and the SNP identification often converted to chromosome and position. It is important to mention that these files have no headers and each time manual data manipulation is necessary, the first row always contains sample information. Along with these three file formats, a .log file is produced reporting the commands of each step taken and possible warnings or errors. Finally, the .hh format file saves heterozygous haploid warnings. For more information about the column structure of each file visit the official PLINK (http://zzz.bwh.harvard.edu/plink/plink2.shtml) documentation.

2.2.3 Autosomal regions and coding in PLINK

Pseudo-autosomal regions (PAR1 and PAR2) of the X and Y chromosomes pair and recombine during meiosis like autosomes, but the recombination activity in PAR 1 is different between sexes (A. Flaquer et al., 2008). PAR 1 is located at the terminal region of the short arms and spans the first 2.7Mb (rough estimate) of the proximal arm of the human sex chromosomes. On the other hand, PAR2 encompasses the distal 320kb of the long arm of each sex chromosomes (D J. Cotter 2016), resulted from at least two duplications from the X chromosome to the terminal end of the Y chromosome(Charchar et al., 2003). To date, at least 24 expressed genes have been identified in the PAR1 region (A Helena Mangs and Brian J Morris, 2007) and at least 5 in PAR2 region (Rüdiger Jörg, Blaschke Gudrun Rappold., 2016). Table 3 demonstrates the coordinates available from the Genome Reference Consortium (`https://www.ncbi.nlm.nih.gov/grc`). Autosome region codes in PLINK are considered from 1 through 22. The X and the Y chromosome are annotated on 23 and 24 respectively, while X chromosomes pseudo-autosomal region is annotated as a separate 'XY chromosome' which has coding number 25. PLINK annotates by default on coding number 26 the Mitochondrial (MT) codes.

Name	Chromosome	Start	Stop
PAR 1	X	60,001	2,699,520
PAR 2	X	154,931,044	155,260,560
PAR 1	Y	10,001	2,649,520
PAR 1	Y	59,034,050	59,363,566

Table 3: Physical locations of pseudo-autosomal and X-translocation regions, coordinates from the Genome Reference Consortium for GRCh37.p13,`https://www.ncbi.nlm.nih.gov/grc/human`

2.2.4 Sex ID, paternal and maternal ID

The value of -9 is considered from PLINK as ''missing", so we recode the 3rd and 4rth column of the .fam file (which corresponds to paternal, maternal id respectively) to 0, considered from PLINK as "not-provided" and causes no problems while running the commands. In Addition, we examine the 5th column of the .fam file, containing the sex information of the under investigation sample. If missing values are present, we examine the manifest of each batch to assign 1 for men and 2 for women. Saving the original files before making major changes on the core files is a common technique used downstream multiple times.

2.2.5 Fluidigm annotation

Fluidigm Corporation (https://www.fluidigm.com/) provides sample identity and integrity checks of genomic data, by validating variant call reproducibility. The combination of the Fluidigm and the 79000HT Sequence Detection system reassures high concordance rates and visual overlay of clusters from dynamic arrays. SNP Genotyping concordance is defined as an allele call that is identical to the previously determined allele type for the sample (F.Yamatomo el at., 2015). In the analyzed datasets, only 10 SNPs have been prior changed to match Fluidigm id documentation. There was a need to convert these SNPs back to the initial SNP identification system in order to have consistency across the multiple under investigation datasets.

2.2.6 Genome reference and strand update

It is essential to pay attention to the build of each exome type and if required, to remap the build in order to have the same reference assembly and strand orientation across the combined datasets. In this study, all genotyping chips were on genome build GRCh37. Illumina genotype calls are not by default using the TOP (forward) strand. We had to update the chromosome, position and strand of the binary files with the shell script developed by Neil Robertson in combination with the corresponding array version for each dataset respectively (`https://www.well.ox.ac.uk/~wrayner/strand/index.html`).

2.3 Considerations when merging cohorts from different array versions

One of the most computationally and programmatically demanding challenges of this master thesis is that we had the scenario of merging four different data-sets where the three of them were from different genotyping array versions (table1). Because of the complicated genetic assumptions of Illumina genotyping data, special handling of the initial files needs to be made before processing to the QC steps. Illumina's exome arrays focus on exonic variants which represent less than 2% of the genome, but contains approximately 85% of know disease-related variants (Dijl EL et al., 2012), making this method in combination with imputation a cost-effective alternative to whole-genome sequencing. In this chapter, we describe the methodology used to overcome the challenges of QC when merging data of different array types and produce a high-quality merged data set for different kind of genetic analyses.

2.3.1 Creation of additional files contain probe sequences and Ids

Before we started any data manipulation, we have created two files. The first one is called "Id listing" having 590,623 official Illumina ids from all arrays involved and how many times each id was present for each array respectively (table 4). The second file is called "Probe Listing" containing 553,860 probe sequences from all arrays involved and how many times each probe sequence is present across the 3 different arrays (table 5). The idea is to compare those files and identify possible mismatches in terms of variants' counts and identification.

Variant id and Chr:pos	Count	Illumina array version		
		-12v1-1_A	-24v1-1_A	-24v1-3_A1
rs2566-131_T_R_1891342930	3	rs2566	rs2566	rs2566

Table 4: IDListing.txt file contains 553.860 variant ids along with the chromosome position of each array and the number of times these were counted across the 3 different array versions.

Probe Sequence	Count	Illumina array version		
		-12v1-1_A	-24v1-1_A	-24v1-3_A1
GTCAACAGCAGAGTGTGTATAGCTGTCAACAAAACGCTAAAACCACAGG	3	exm828933	exm828933	exm828933

Table 5: ProbeListing.txt file contains 590,623 probe sequences of all arrays and the number of times a probe sequence was counted across the 3 different arrays versions.

2.3.2 Complications and solutions

The most common way to accurately handle possible issues is by comparing the two files and recoding the under investigation datasets per chromosome position. This step takes place to extract the overlapping variants among the three arrays' versions. We noticed the following:

1. There were cases in the probe listing file where even if the count was 3, the variant id was missing in at least one array due to duplication in one of the other array visions (table 6).

2. Cases where a variant was mention with different identification name across the array versions (table 7).

3. We found variants with the same identification reported in a different chromosome position, with 1 base pair difference (table 8).

4. We found variants with different identification and having the same chromosome position (table 7).

Probe Sequence	Count	Illumina array version		
		-12v1-1_A	-24v1-1_A	-24v1-3_A1
TTGAGCACTCTTCTGTAATCTCATGA GGTGTCCAGGGAAGAGACAATGAT	3	rs1873983	rs1873983	

Table 6: Variants' count issue.

Rs ids	Morgans	Base Pair	Allele A	Allele B
exm2253593	0	900427	A	G
exm596	0	900427	A	G

Table 7: Same chromosome position with different variant id.

Rs ids	Morgans	Base Pair	Allele A	Allele B
indel.20904	0	102650052	I	D
indel.20904	0	102650053	I	D

Table 8: Same variant id in a different chromosome position.

In order to overcome the above-mentioned complications, and make sure that each variant is present across the arrays, we extracted from the "ProbeListing.txt" file only the variants having count 3 with the condition of not having null values in any of the other arrays' versions, saved as "Finallistwith3.txt". Then we joined the ids and chromosome positions in all four .bim files and saved them with a new file named "merged.bims.txt". To create the file "inclusionlist.txt" we extracted from the "Finallistwith3.txt" all the variants that have a different position in any of the "merged.bims.txt" file. Finally, we extracted the "inclusionlist.txt" file from each original dataset-cohort respectively, to keep only the overlapping variants across the three arrays' versions. Before merging the four different cohorts with PLINK, all multi-allelic variants (saved as "multialleliclist.txt") have been removed.

Chapter 3

Findings and Results

3.1 Sample QC

3.1.1 Overview

Briefly, as we mentioned in chapter I, genotypes were called using Illumina Gencall application and mapped to GRCh37. We handled each cohort separately for updating the paternal and the maternal ids (Cohorts 1, 2, 3, 4), reversing the Fluidigm annotation to the original rs ids (Cohorts 2, 3) and removed any sample that was not part of the study design. Lastly, we updated the strand separately for each cohort respectively per array version. After overcoming the technical challenges mentioned in section 2.4.3, we have generated a final merged PLINK binary file, containing 520.123 SNPs and 220 individuals.

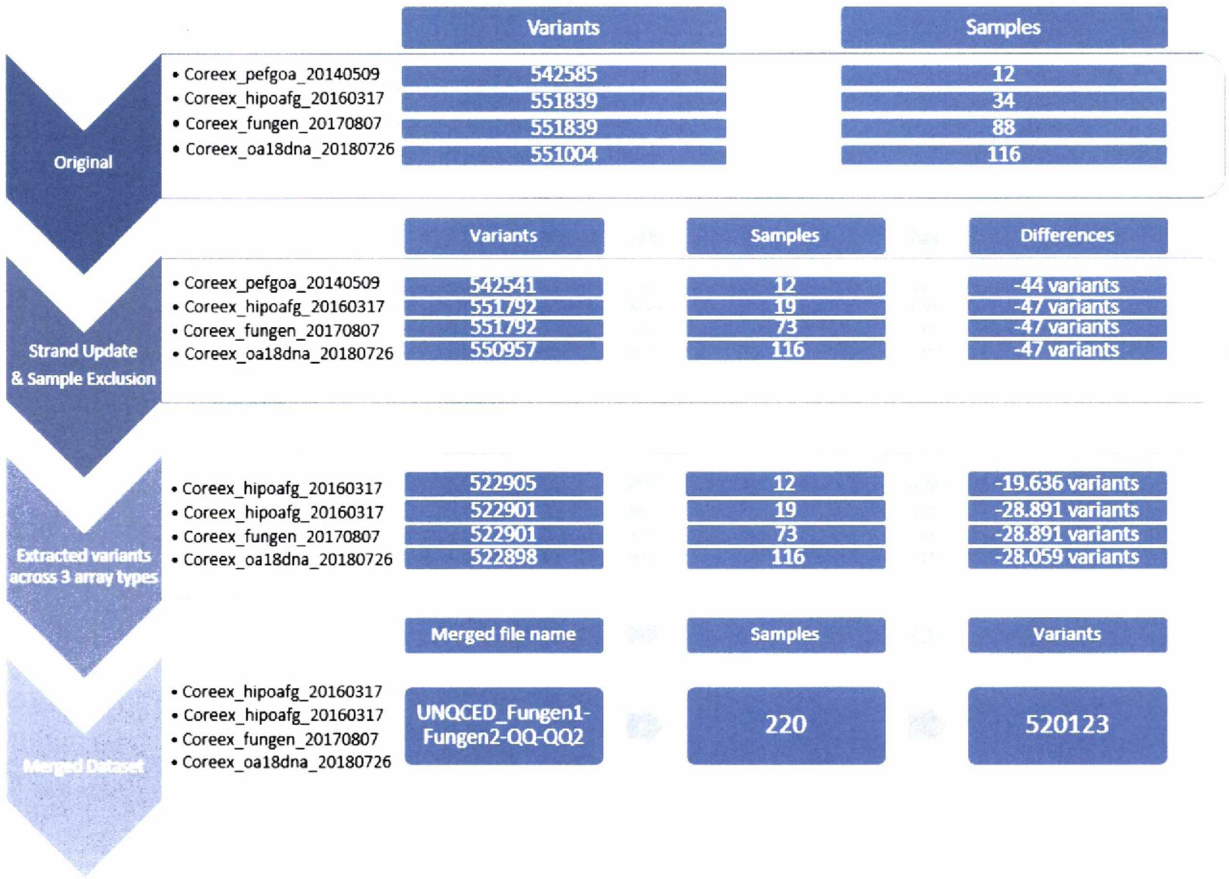

Figure 4: An overview of the workflow until merging four cohorts in one, providing information about the variant and sample progression of each step taken denoted as an arrow on the left-hand side.

3.1.2 Pre-filtered QC results

Before the actual QC, we have applied on the pre-filtered merged dataset a threshold of 90%, were no samples having a call rate<90% excluded (Figure 5). The file to be analysed is ".imiss" for individual missingness (Sample QC) and ".lmiss" for locus missingness (SNP QC). In both cases, no individuals were excluded. caption subcaption

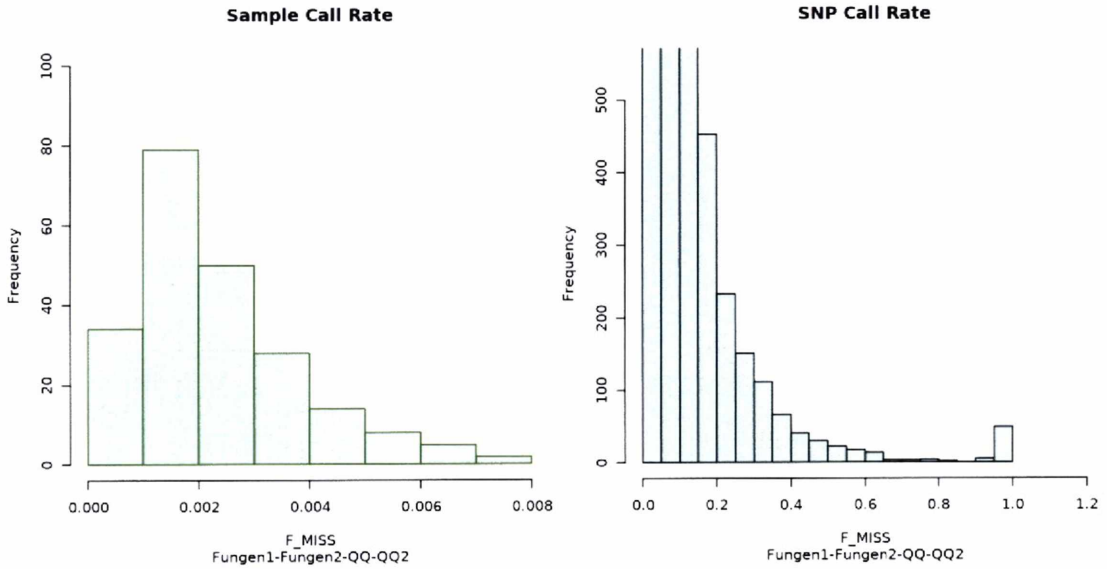

Figure 5: Apply 90% call rate threshold to remove bad quality genotyped samples and SNPs.

3.1.3 Sample call rate

Samples were pre-filtered and excluded with call rate<98%. At this stage, no individuals were removed and no significant missingness reported from a genotyping calling perspective (Figure 5).

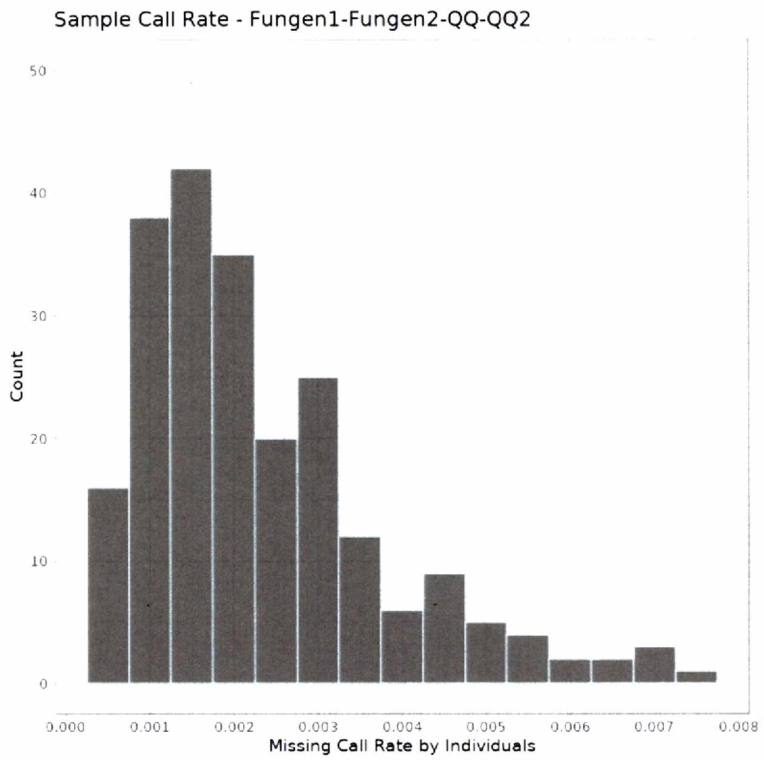

Figure 6: Individual 98% Call rate of the merged cohort data set. On the y-axis are the frequency rates, while at x-axis the missingness rates.

3.1.4 Sex discrepancies

We did not identify any gender mismatching between the one revealed from genotyping data and the one reported during the sample collection (Figure 7). The X chromosome inbreeding estimates were as expected close to 1 and 0 for males and females respectively. The color represents the self-reported gender match of each individual while each dot represents the actual gender identification from the DNA genotyping process. The regression lines represent every single line that best fits on each group (males and females). In total, there are no individuals to be excluded.

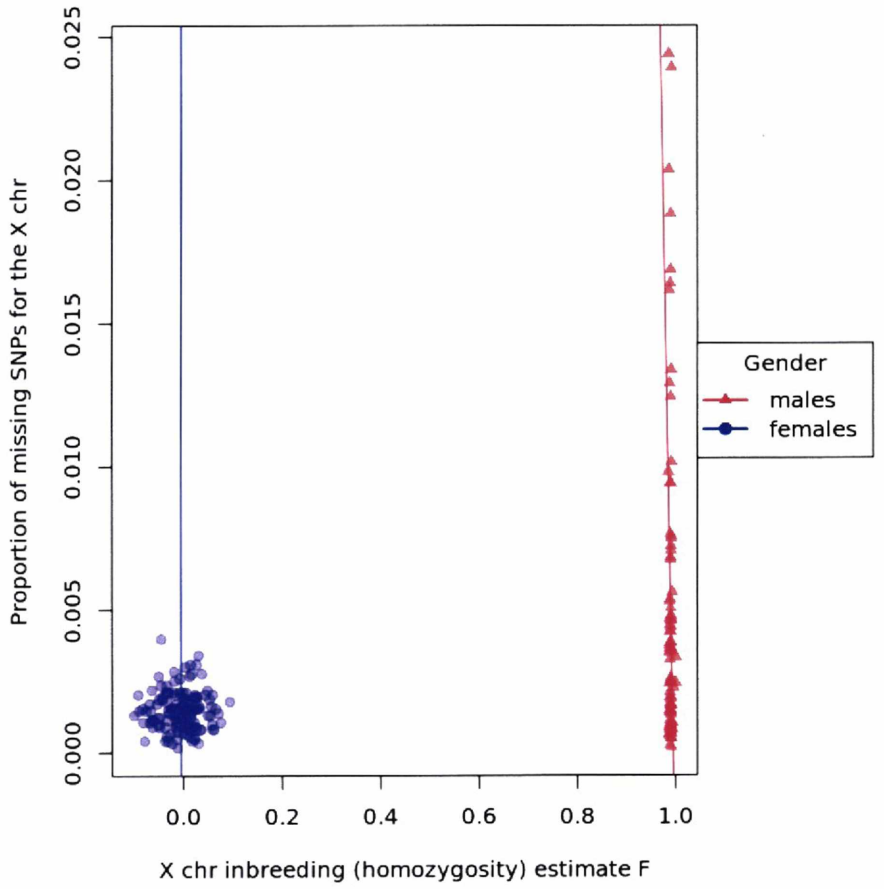

Figure 7: Identifying sex mismatches, while plotting in the y-axis the proportion of missing SNPs for the chromosome X and in x-axis the probability an individual to have two identical alleles from a single ancestor.

3.1.5 Heterozygosity estimation

We plot the autosomal heterozygosity rates separately for MAF1% (figure 8) and MAF<1% (figure 10) to separate the common with the rare variants (MAF<1%). The MAF bin is useful as rare genotypes have different characteristics from the common ones (e.g. less power to detect an association and the genotype calls will be less certain (Sally R. Ellingson, David W. Fardo3., 2016)). In total, three individuals are excluded, two samples for MAF1% (figure 9) and another one for MAF<1% (Figure 10), having 3 standard deviations away from the mean as the threshold.

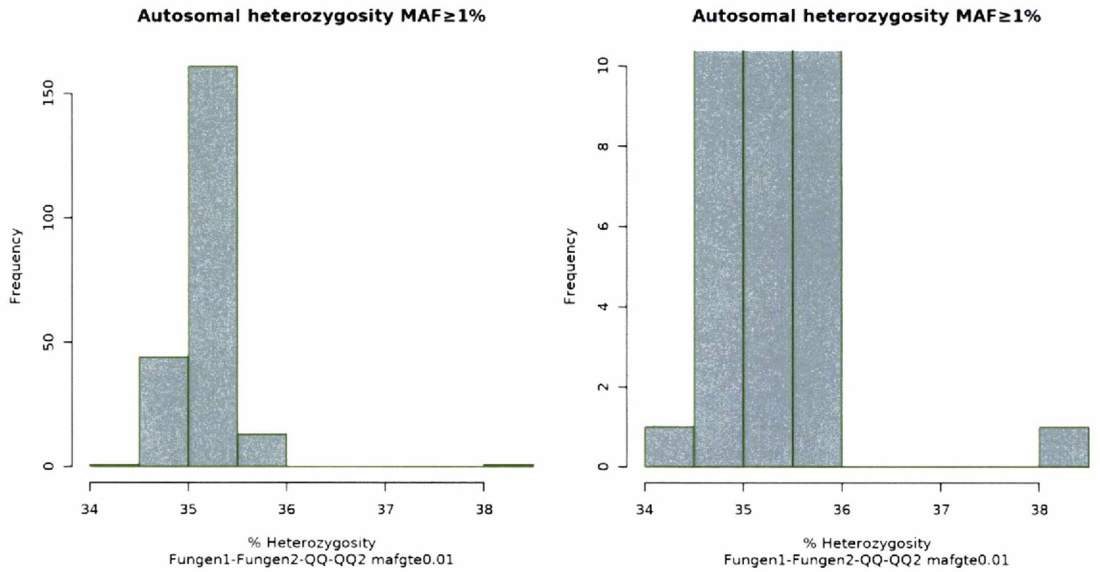

Figure 8: Autosomal heterozygosity rates for MAF1%. The left plot shows how the samples are distributed based on heterozygosity rates across the complete merged dataset, while on the right one we zoom in the tails of the histogram.

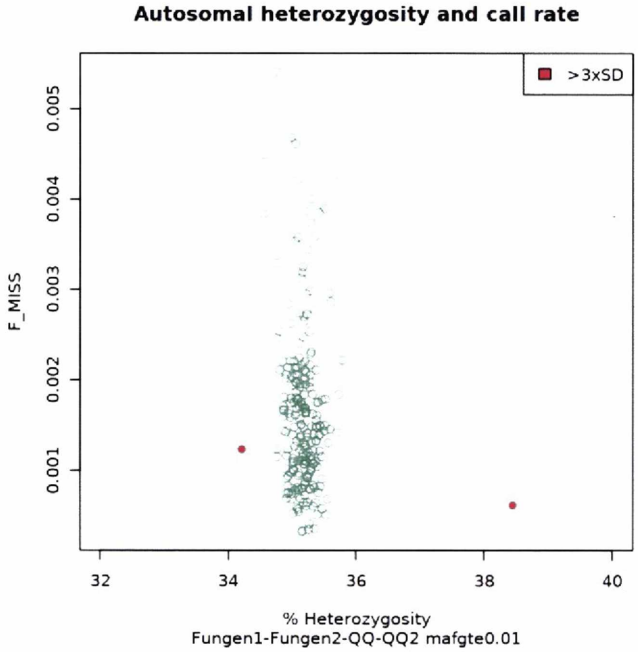

Figure 9: Percentage of autosomal heterozygosity (x-axis) against rates of missingness (y-axis). Each dot represents an individual. The red dots indicate individuals with 3-sd away from the heterozygosity mean.

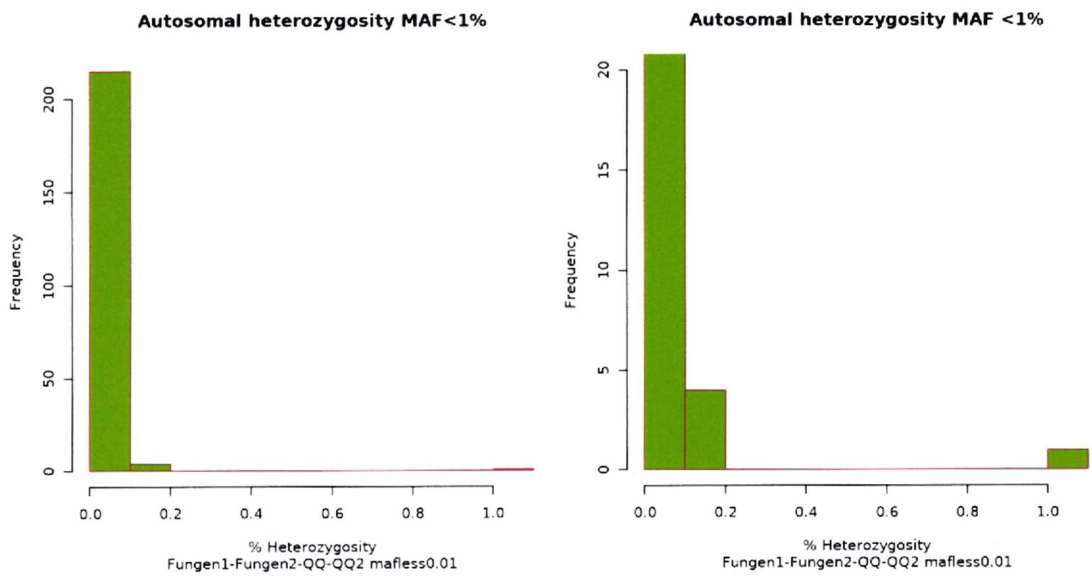

Figure 10: Autosomal heterozygosity rates for MAF<1%. The left side shows how the samples are distributed based on heterozygosity rates across the complete merged dataset, while on the right one we zoom in the tails of the histogram.

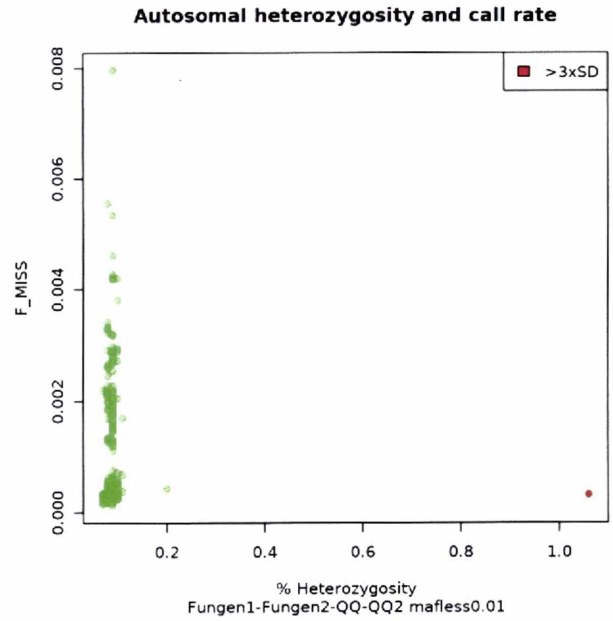

Figure 11: Percentage of autosomal heterozygosity (x-axis) against rates of missingness (y-axis). Each dot represents an individual. The red dots indicate individuals with 3-sd away from the heterozygosity mean.

3.1.6 Duplicates and relatedness

In our analysis, we kept individuals having less than 2nd degree of relatedness (pihat<0.2). Individuals with pihat>0.9 are either duplicates or identical twins, where (0.2<pihat<0.9) indicates related pairs (Figure 12). Five individuals are excluded in total, three pairs of them are duplicates or monozygotic twins, one pair is 1rst degree relatives and one pair is 2nd degree relatives.

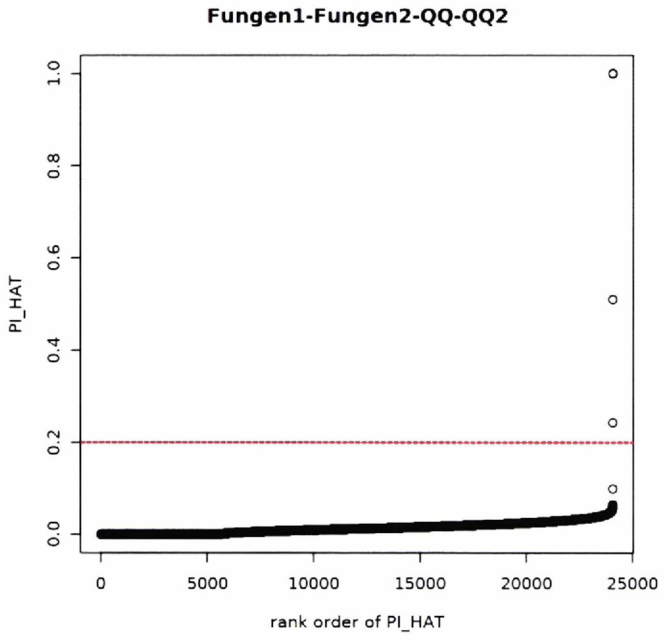

Figure 12: Ranking the pihat values to identify related and duplicated/twins individuals.The red line demonstrates the exclusion threshold (pihat>0.2).

Generally, when finding a pair of related individuals, a decision needs to be made between which one should be removed. This decision is usually based on the following criteria:

1. Based on the call rate between the duplicated or ancestry related individuals, we want to keep the one with the highest call rate.
2. If the sample call rate is the same:
    - The decision can be based on disease status. It is commonly used to keep cases against controls.
    - Individuals that are related to more than one participants are being removed.

Duplicate pairs along with missingness rates

Sample ID 1	Sample ID 2	Missingness 1	Missingness 2
urn:wtsi:433501_A04_OAfg6308476	urn:wtsi:493153_G01_OAfg6308476	0.0005096	0.002091
urn:wtsi:433501_G02_OAfg6308466	urn:wtsi:493153_E01_OAfg6308466	0.0003628	0.001834
urn:wtsi:433501_G03_OAfg6308474	urn:wtsi:493153_F01_OAfg6308474	0.0003302	0.001296
urn:wtsi:493153_H10_fungenQQ6828557	urn:wtsi:533231_E04_fungenQQ7246857	0.002658	0.0002731
urn:wtsi:533231_D01_fungenQQ7246832	urn:wtsi:533232_H01_fungenQQ7442739	0.001054	0.001944

Table 9: Duplicate pairs of individuals along with the missingness rates.

3.1.7 Population structure and stratification

The most common method used during population stratification is to merge the under-investigation dataset with the 1KG and run a PCA analysis to produce a multidimensional reduction plot and use it as a decision tool (Figure 13). An important notice is that we kept only the overlapping variants between our file and the 1KG so we have an equal representable amount of SNPs to compare. One sample is excluded at this stage, and when zooming in further on the PCA plot we identify another three exclusion targets (Figure 14). In order to make a subjective decision, we could either use a clustering algorithm or since the under investigation population is British, we have used as representative reference population the UK Household longitudinal study (https://www.understandingsociety.ac.uk/) – indicated with green colour (Figure 15), which is one of the largest panel surveys in the world, supporting social and economic research. Its sample size is 40.000 households from the United Kingdom with approximately 100.000 individuals. To sum up, during this stage three individuals are to be excluded.

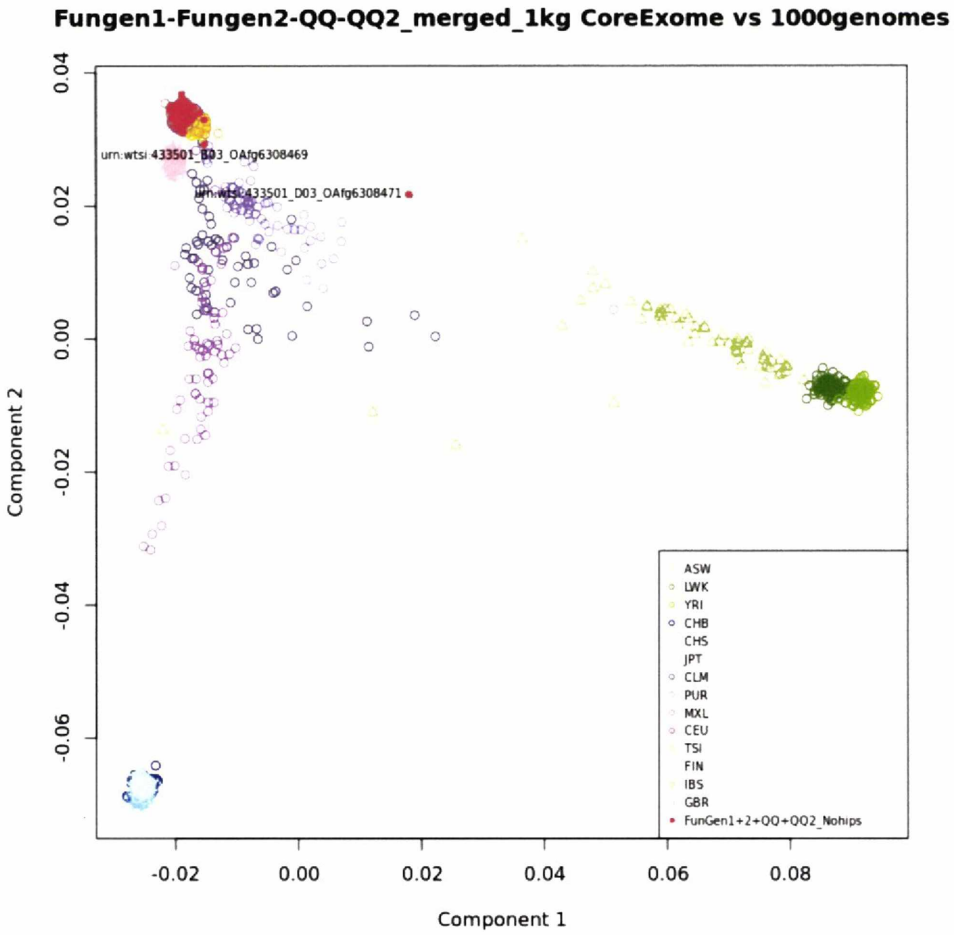

Figure 13: Multidimensional Scaling (MDS) and PCA of the four cohorts combined with populations from the 1000 Genome Project. Individuals from the FunGen project are depicted by red solid circles and clustered close to the other European-ancestry individuals.Abbreviations: ASW, Americans of African ancestry in south-western USA; LWK, Luhya in Webuye, Kenya; YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan; CLM, Colombians from Medellin, Colombia; PUR, Puerto Ricans from Puerto Rico; MXL, Mexican ancestry from Los Angeles, USA; CEU, Utah residents with Northern and Western European ancestry; TSI, Toscani in Italy; FIN, Finnish in Finland; IBS, Iberian population in Spain and GBR, British in England and Scotland.

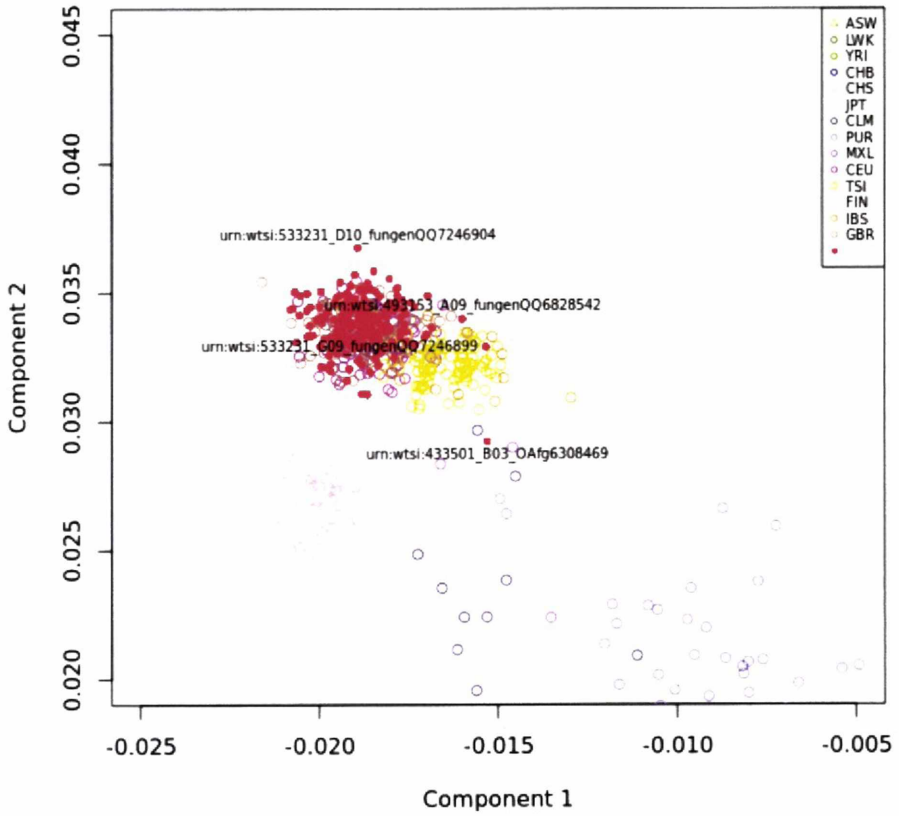

Figure 14: MDS zooming in the population of interest, depicted by red solid circles.Abbreviations: ASW, Americans of African ancestry in southwestern USA; LWK, Luhya in Webuye, Kenya; YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan; CLM, Colombians from Medellin, Colombia; PUR, Puerto Ricans from Puerto Rico; MXL, Mexican ancestry from Los Angeles, USA; CEU, Utah residents with Northern and Western European ancestry; TSI, Toscani in Italy; FIN, Finnish in Finland; IBS, Iberian population in Spain and GBR, British in England and Scotland.

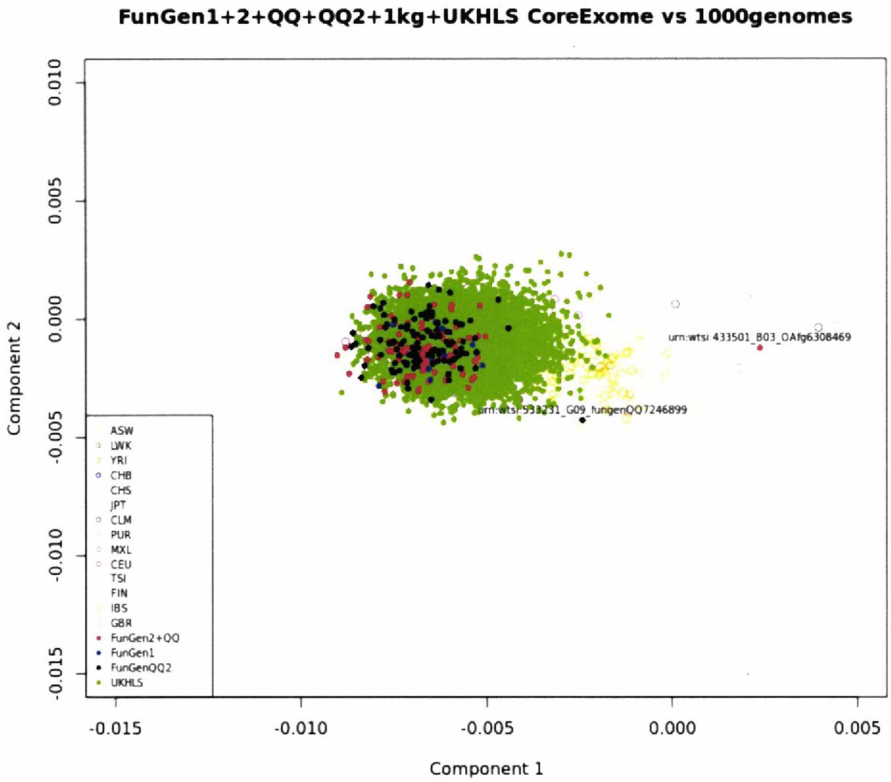

Figure 15: Multidimensional Scaling (MDS) and PCA of the combined with UKHLS cohort, depicted with green solid circles and split the under investigation dataset in its initial form (four cohorts; fungen1 depicted with dark blue solid circles; Fungen2 and FunGen QQ2 with red solid circles; FunGen QQ2 depicted with black solid circles) .Abbreviations: ASW, Americans of African ancestry in south-western USA; LWK, Luhya in Webuye, Kenya; YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan; CLM, Colombians from Medellin, Colombia; PUR, Puerto Ricans from Puerto Rico; MXL, Mexican ancestry from Los Angeles, USA; CEU, Utah residents with Northern and Western European ancestry; TSI, Toscani in Italy; FIN, Finnish in Finland; IBS, Iberian population in Spain and GBR, British in England and Scotland.

3.1.8 Summary of individual-based QC

During the individual-based QC, samples were pre-filtered and excluded with a call rate $< 90\%$. Samples were checked for gender discrepancies, excess heterozygosity at MAF1% and MAF<1%, duplicates, relatedness and ethnicity, leading to 9 excluded individuals and 211 samples to proceed for SNP QC (table 10).

FunGen Merged Data set	Number of variants
Number of samples – all cases	237
Samples for Quality Control	220
Sample Call Rate < 98%	0
Gender Mismatches	0
Heterozygosity (MAF ≥1%)	2
Heterozygosity (MAF <1%)	1
Duplicates (pi-hat > 0.9)	3
Relatedness (pi-hat > 0.2)	2
Ethnicity Outliers	3
Total number of samples to exclude	11 (9 unique)
Total number of samples to include proceeding for SNP QC	211

Table 10: Summary of the individual-based QC exclusions. QC, Quality Control, MAF, Minor Allele Frequency.

3.2 SNP Based Quality Control

3.2.1 Overview

Before started the marker QC, all individuals that did not pass the criteria of the individual-based QC are excluded. We performed the SPN-based QC separately for the autosomal region; the non-pseudo-autosomal (PAR) of the females (coded as chromosome 23 in PLINK) and the PAR 1 and PAR 2 (coded as chromosome 25 in PLINK) of both males and females. We apply a 98% SNP call rate filter, where all SNPs that have more than 2% missingness were excluded. An additional filter for the SNP QC is to apply the HWE principle where we generated HWE pvalues and a QQ plot of the log-p-values of the SNPs for the under investigation samples. We excluded all variants having a HWE p-value $<$ 1e-4. Last but not least, we excluded all non-pseudo-autosomal SNPs that have haploid male genotypes – automatically saved by PLINK on the ".hh" file format (Figure 16).

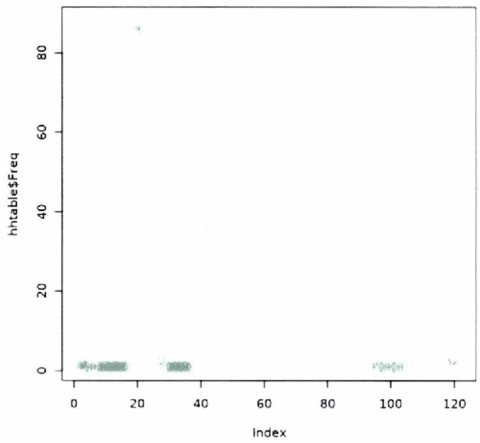

Figure 16: Amount of SNPs containing heterozygous haploids.

3.2.2 Summary of SNP QC

Variant Filtering	FunGen Collection		
	Male / Females	Females	Males / Females
	Autosomal (1-22)	X-nonPAR (23)	X-PAR 1 &PAR 2 (25)
Individuals for QC	222	122	222
Variants in data	505,307	12,397	155
Call rate 98%	9906	283	1
HWE	85	0	0
Unique SNPs	42	283	1
Total SNP exclusion	9,948	283	1
Total SNP inclusion	495,359	12,114	154
Total amount of SNPs with heterozygous haploid males		122	
Total amount of SNPs proceeding for imputation		***509,769***	

Table 11: Summary of the SNP-based QC exclusions. QC, Quality Control, HWE, Hardy Weinberg Equilibrium, PAR, Pseudoautosomal.

3.3 Imputation using the Haplotype Reference Consortium - HRC

3.3.1 Preparation for imputation on Michigan server

After removing 9 individuals and 10,354 SNPs that did not pass the QC criteria, we have proceeded with the cleaned dataset to impute the missing variants, using the Haplotype Reference Consortium v1.1 (hg19) as reference panel on the Michigan imputation server. First, we have used the HRC preparation checking tool by Will Rayner https://www.well.ox.ac.uk/ wrayner/tools/HRC-1000G-check-bim-v4.2.9.zip, which identifies if the strand, alleles, positions and Ref/Alt alignments and frequencies are consistent between the HRC panel and our SNP files. Moreover, this tool automatically produces a set of PLINK commands, and based on the results of this checklist, it removes or updates SNPs that are not consistent. As an additional quality check, we have used the checkVCF tool (`http://qbrc.swmed.edu/zhanxw/software/checkVCF/checkVCF-20140116.tar.gz`), which includes the checkVCF.py script, the reference genome (hs37d5.fa) in FASTA format and an index file (hs37d5.fa.fai). The aim of the last step is to detect any mismatching on the reference alleles with the reference genome. An important notice is that the variant identification of the PLINK files were per chromosome and position rather than the exome array identification.

3.3.2 Imputation results

After completing the pre-imputation QC steps, we have uploaded the files on VCF format on the Michigan imputation server. Our dataset was consisted of 211 individuals and 281231 autosomal SNPs (chromosome 1 to 22). From the 381,231 uploaded SNPs, the total remaining sites and the number of SNPs with reference overlap were 276,127, and no allele switch or strand flip was necessary. In addition, 36 potential frequency mismatches identified. After completing the genotype imputation process, using the Minimac 4 (1.2.4) algorithm and Eagle for phasing, the final output consist of 39,127,678 SNPs (Table 12). These SNPs are distributed with different alternative allele frequencies on the genome-wide level.

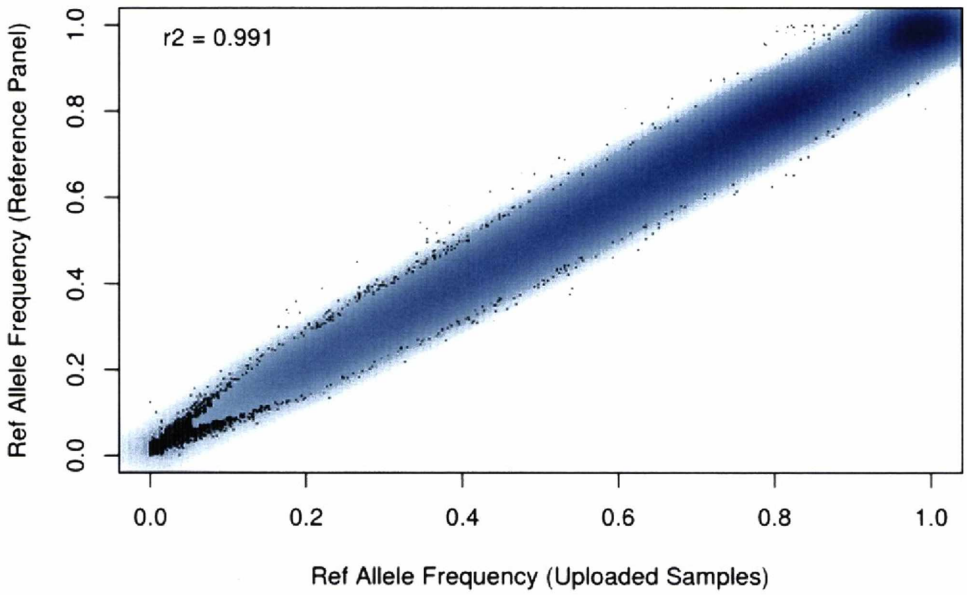

Figure 17: Allele-Frequency Correlation after imputation, showing the densities of the frequencies falling into each part. The first 5000 points from the areas of the lowest regional densities are plotted. A comparison between the allele frequency of the reference panel (y-axis) with the allele frequency of the uploaded dataset (x-axis).

3.3.3 Post-imputation QC metrics

To ensure genotype imputation quality, we perform a post-imputation QC, considering the INFO score (Table 12), ranging from 0 to1, where higher value represents the increased quality of the imputed variants. Another quality control metric is the r-square (Figure 17), although both metrics shown are not directly comparable (Schurz H et al., 2019).

Imputation INFO score		
INFO score	**Count**	**%**
0.0	26518074	67.773
0.1	679826	1.737
0.2	373712	0.955
0.3	301050	0.769
0.4	327402	0.837
0.5	410320	1.049
0.6	636108	1.626
0.7	1040420	2.659
0.8	1778403	4.545
0.9	7035568	17.981
1.0	26795	0.068
Total	39127678	100.000

Table 12: Imputation's INFO score on Genome-Wide level. Variants were binned by INFO score values.

Alternative Allele Frequency		
Alternative Allele Frequency	**Count**	**%**
0.0	34499098	88.171
0.1	1185497	3.030
0.2	790719	2.021
0.3	599650	1.533
0.4	484305	1.238
0.5	394599	1.008
0.6	338607	0.865
0.7	285349	0.729
0.8	239455	0.612
0.9	309397	0.791
1.0	1002	0.003
Total	39127678	100.000

Table 13: Alternative allele frequency on genome-wide level. Variants binned by alternative allele frequency percentages.

The post-imputation QC checks are provided per chromosome. In the following figures (18, 19, 20, 21), we examine with the same method each chromosome

subsequently, to identify any inconsistencies between the under investigation cohort and the HRC panel. For practical reasons, we demonstrate only chromosome 1, which is the largest in terms of variants' counts. Again, we start by using the INFO Score metric (Table 14) using various plots for visualization like Bar plot (Figure 18), Manhattan plot (Figure 19), and ordering the variants of chromosome 1 per position (Figure 21).

Imputation INFO score on Genome-Wide level

INFO score	Count	%
0.0	2088941	68.0452
0.1	54580	1.7779
0.2	30146	0.9820
0.3	24389	0.7944
0.4	27439	0.8938
0.5	33727	1.0986
0.6	51451	1.6760
0.7	83658	2.7251
0.8	140614	4.5804
0.9	533310	17.3721
1.0	1676	0.0546
Total	3069931	100.000

Table 14: Imputation INFO Score for chromosome 1. Variants binned by INFO score.

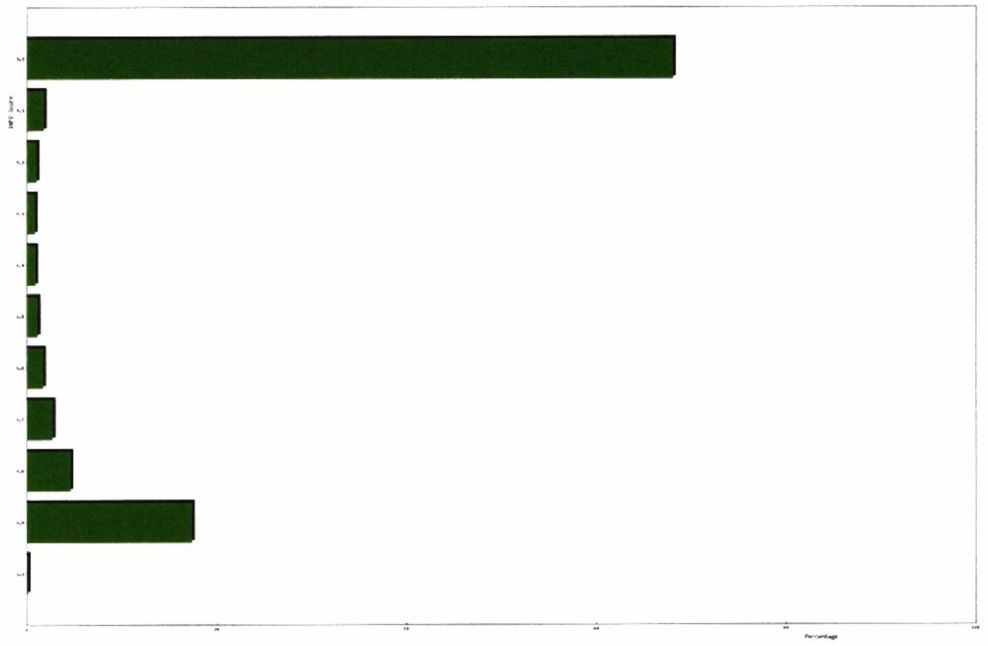

Figure 18: Imputation INFO Score across chromosome 1. Variants binned by INFO score. The vertical U-shape distribution symbolizes that the imputation is robust enough.

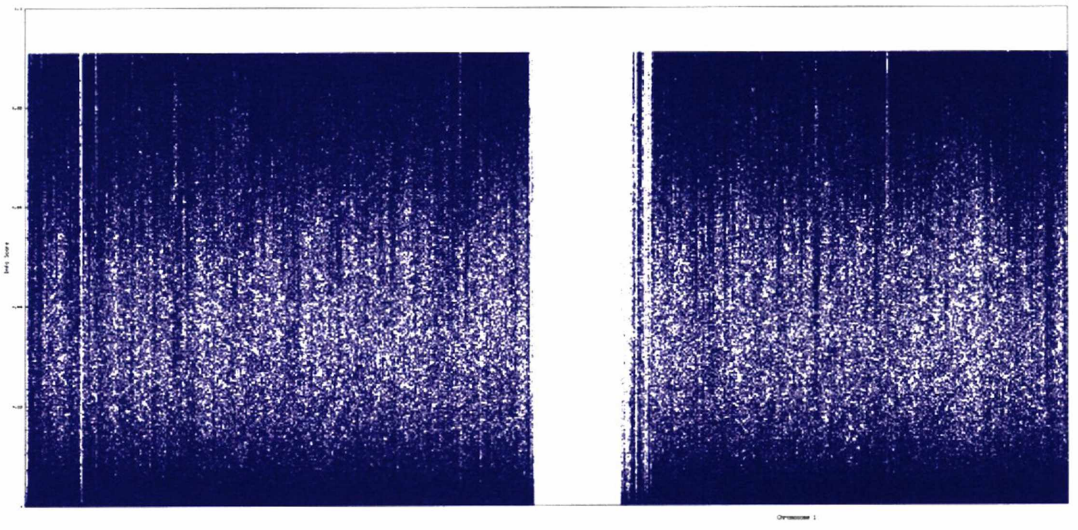

Figure 19: Manhattan Plot of the INFO Score across chromosome 1. Genomic coordinates are displayed along the X-axis, with the negative logarithm of the association P-value for each SNP of chromosome 1 displayed on the Y-axis

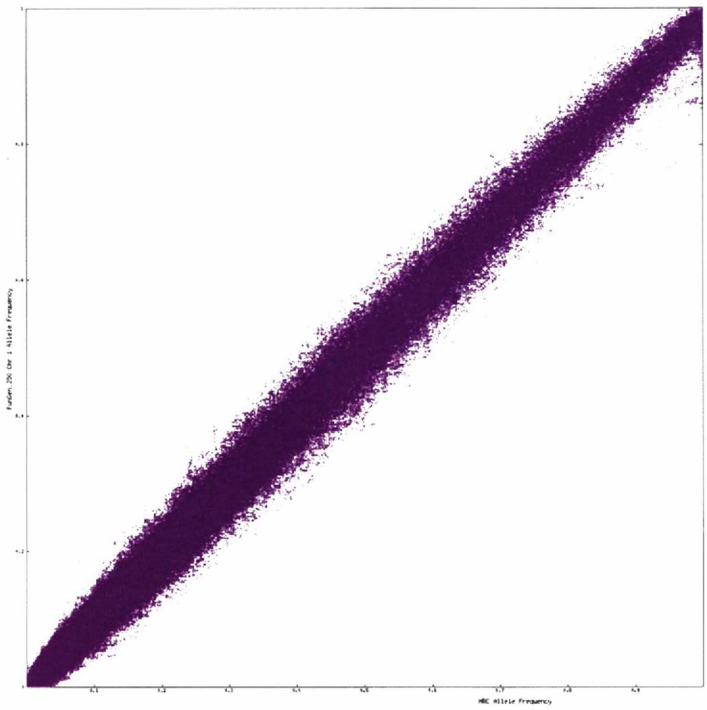

Figure 20: Alternative Allele Frequency of chromosome 1 compared to the HRC reference panel. Variants should be clustered without any extreme outlier values which indicates the quality of the imputation.

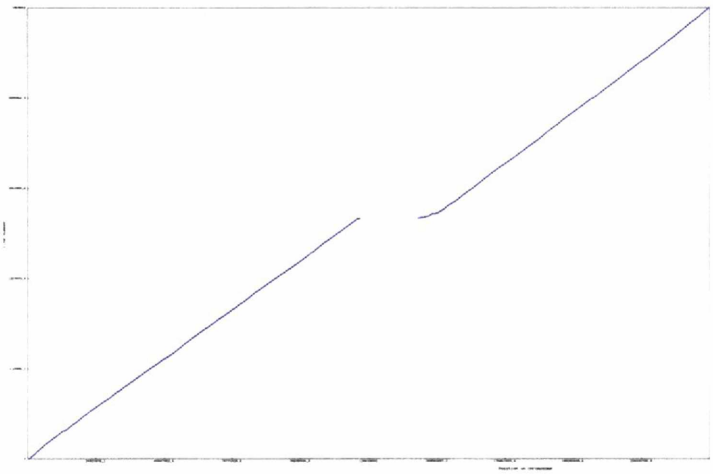

Figure 21: Ordering the variants of chromosome 1 per position to identify any inconsistency. The result indicates powerful imputation process. The central gap refers to the non-coding DNA region of centromere.

Chapter 4

Discussion

During the last years, human has developed various techniques to extract and analyze DNA. Through next-generation sequencing technologies, we have managed to create a huge knowledge base to further revolutionize genomic research. Genotype array technologies provide a cost-effective solution to analyze genomic data. Because these arrays do not directly genotype all variation in the genome, statistical methods like genotype imputation and linking missing values to publicly available reference panels like Hap Map and the 1KG, allow to test whether there is an association between a trait of interest (e.g. a disease) and specific variation (SNPs). In order to use these publicly available tools one of the first and most critical procedures that take place before imputation and any kind of association test is Quality Control (QC) of the under investigation samples. Once a high-quality group of imputed SNPs is available, additional analyses can be carried on such as RNA-seq, expression of quantitative trait loci (eQTLs) and GWAS.

The most essential element before contacting QC is to properly design the experiment. It is critical to avoid variables affecting the genotyping quality which is usually caused during the DNA storage and shipping conditions, extraction during genotyping, and plate effects while processing batches. In addition, when merging different cohorts it would be ideal to have batches from the same array type. If not, it is possible to proceed on merging multiple cohorts through manual manipulation of the provided genotype datasets in combination with the published manifests of the genotype provider (e.g. Illumina) and still produce a high-quality genetic dataset. Due to the complications of this procedure, it is highly advised to considerate the manifest of each batch carefully and examine the technical details of each cohort. Fewer differences in the examined studies in the sample collection and genotyping phases will lead to minimizing issues in the downstream analyses.

Nowadays there are many bioinformatics tools, techniques and packages facilitating direct screening with automatic measurement of an individuals' genotype at thousands of markers. The advantage of the «manual QC» reassure the maximum quality and efficiency in controlling each phase of the process, especially when dealing with technical details such as standardization between multiple cohorts of different genotyping arrays or platforms. A future direction is to not focus only on the autosome regions but also integrate the X and Y chromosomes. Especially for the X chromosome, which contains rich genetic information in terms of population history, the analysis framework is almost completed, including variant imputation in well-know reference panels. There are publicly available bioinformatic tools developed to facilitate the need of special handling of chromosome X such as XWAS mainly focused on Affymetrix arrays (Geo F, et al., 2015) and SHAPEIT (`http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html`). Filling up this gap could reveal an area of research with great findings, especially for traits and diseases with sex-specific features (R. Koenig et al., 2013).

In our study, we have covered a number of computational and statistical challenges arising from the QC and interpretation of GWAS data. We have used the most recently developed techniques for QC procedures of genotype and impute data and we presented a protocol for keeping to the minimum type I and type II errors in the downstream genetic analyses. In addition, we have reduced the likelihood of failed replication of the study by minimizing the biases in the under investigation samples and provided the exact command used as a tutorial package (`https://rpubs.com/fondan/quality_control_of_genotype_data`) along with this thesis.

It is conceivable that in the future we will be flooded with more data in GWAS studies and there are significant challenges in the processing and quality control of common rare and novel variants. There is much to learn from many on-going activities in statistical applications and dimension reduction methods. At the same time, there is no question that GWAS data will motivate new approaches from multidisciplinary areas in the near future.

Bibliography

[Altshuler, Daly LanderAltshuler .2008] altshuler2008geneticAltshuler, D., Daly, MJ. Lander, ES. 2008. Genetic mapping in human disease Genetic mapping in human disease. science3225903881–888.

[Anderson .Anderson .2010] anderson2010dataAnderson, CA., Pettersson, FH., Clarke, GM., Cardon, LR., Morris, AP. Zondervan, KT. 2010. Data quality control in genetic case-control association studies Data quality control in genetic case-control association studies. Nature protocols591564.

[BL. Browning BrowningBL. Browning Browning2009] browning2009unifiedBrowning, BL. Browning, SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics842210–223.

[SR. Browning BrowningSR. Browning Browning2011] browning2011haplotypeBrowning, SR. Browning, BL. 2011. Haplotype phasing: existing methods and new developments Haplotype phasing: existing methods and new developments. Nature Reviews Genetics1210703–714.

[Consortium .Consortium .2012] Consortium, GP. . 2012. An integrated map of genetic variation from 1,092 human genomes An integrated map of genetic variation from 1,092 human genomes. Nature491742256.

[Duncan BrownDuncan Brown2018] duncan2018genomeDuncan, EL. Brown, MA. 2018. Genome-wide association studies Genome-wide association studies. Genetics of Bone Biology and Skeletal Disease Genetics of bone biology and skeletal disease (33–41). Elsevier.

[DurbinDurbin2014] durbin2014efficientDurbin, R. 2014. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT) Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). Bioinformatics3091266–1272.

[Ellingson FardoEllingson Fardo2016] ellingson2016automatedEllingson, SR. Fardo, DW. 2016. Automated quality control for genome wide association studies Automated quality control for genome wide association studies. F1000Research5.

[Fachal .Fachal .2020] fachal2020fineFachal, L., Aschard, H., Beesley, J., Barnes, DR., Allen, J., Kar, S.others 2020. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nature Genetics1–18.

[Fardo, Ionita-Laza LangeFardo .2009] fardo2009qualityFardo, DW., Ionita-Laza, I. Lange, C. 2009. On quality control measures in genome-wide association studies: a test to assess the genotyping quality of individual probands in family-based association studies and an application to the HapMap data On quality control measures in genome-wide association studies: a test to assess the genotyping quality of individual probands in family-based association studies and an application to the hapmap data. PLoS genetics57.

[Gao .Gao .2015] gao2015xwasGao, F., Chang, D., Biddanda, A., Ma, L., Guo, Y., Zhou, Z. Keinan, A. 2015. XWAS: a software toolset for genetic data analysis and association studies of the X chromosome Xwas: a software toolset for genetic data analysis and association studies of the x chromosome. Journal of Heredity1065666–671.

[Gilissen, Hoischen, Brunner VeltmanGilissen .2012] gilissen2012diseaseGilissen, C., Hoischen, A., Brunner, HG. Veltman, JA. 2012. Disease gene identification strategies for exome sequencing Disease gene identification strategies for exome sequencing. European Journal of Human Genetics205490–497.

[Horwitz, Lam, Chen, Xia LiuHorwitz .2019] horwitz2019decadeHorwitz, T., Lam, K., Chen, Y., Xia, Y. Liu, C. 2019. A decade in psychiatric GWAS research A decade in psychiatric gwas research. Molecular psychiatry243378–389.

[Jeong .Jeong .2017] jeong2017comorbiditiesJeong, H., Baek, SY., Kim, SW., Eun, YH., Kim, IY., Lee, J.Cha, HS. 2017. Comorbidities and health-related quality of life in Koreans with knee osteoarthritis: Data from the Korean National Health and Nutrition Examination Survey (KNHANES) Comorbidities and health-related quality of life in koreans with knee osteoarthritis: Data from the korean national health and nutrition examination survey (knhanes). PloS one1210.

[Jun .Jun .2012] jun2012detectingJun, G., Flickinger, M., Hetrick, KN., Romm, JM., Doheny, KF., Abecasis, GR.Kang, HM. 2012. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data Detecting and estimating contamination of human dna samples in sequencing and array-based genotype data. The American Journal of Human Genetics915839–848.

[Kang, Qin, Niu LiuKang .2004] kang2004incorporatingKang, H., Qin, ZS., Niu, T. Liu, JS. 2004. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. The American Journal of Human Genetics743495–510.

[Kong .Kong .2008] kong2008detectionKong, A., Masson, G., Frigge, ML., Gylfason, A., Zusmanovich, P., Thorleifsson, G.others 2008. Detection of sharing by descent, long-range phasing and haplotype imputation Detection of sharing by descent, long-range phasing and haplotype imputation. Nature genetics4091068.

[König, Loley, Erdmann ZieglerKönig .2014] konig2014includeKönig, IR., Loley, C., Erdmann, J. Ziegler, A. 2014. How to include chromosome X in your genome-wide association study How to include chromosome x in your genome-wide association study. Genetic epidemiology38297–103.

[MI. McCarthyMI. McCarthy2008] mccarthy2008abecasisMcCarthy, MI. 2008. Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN Abecasis gr, cardon lr, goldstein db, little j, ioannidis jp, hirschhorn jn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet9356–369.

[MI. McCarthy .MI. McCarthy .2008] mccarthy2008genomeMcCarthy, MI., Abecasis, GR., Cardon, LR., Goldstein, DB., Little, J., Ioannidis, JP. Hirschhorn, JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics95356–369.

[S. McCarthy .S. McCarthy .2016] mccarthy2016referenceMcCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, AR., Teumer, A.others 2016. A reference panel of 64,976 haplotypes for genotype imputation A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics48101279–1283.

[Mills RahalMills Rahal2019] mills2019scientometricMills, MC. Rahal, C. 2019. A scientometric review of genome-wide association studies A scientometric review of genome-wide association studies. Communications biology211–11.

[Patron, Serra-Cayuela, Han, Li WishartPatron .2019] patron2019assessingPatron, J., Serra-Cayuela, A., Han, B., Li, C. Wishart, DS. 2019. Assessing the performance of genome-wide association studies for predicting disease risk Assessing the performance of genome-wide association studies for predicting disease risk. PLOS one1412.

[Purcell .Purcell .2007] purcell2007plinkPurcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, MA., Bender, D.others 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses Plink: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics813559–575.

[Rahman, Kopec, Anis, Cibere GoldsmithRahman .2013] rahman2013riskRahman, MM., Kopec, JA., Anis, AH., Cibere, J. Goldsmith, CH. 2013. Risk of cardiovascular disease in patients with osteoarthritis: a prospective longitudinal study Risk of cardiovascular disease in patients with osteoarthritis: a prospective longitudinal study. Arthritis care & research65121951–1958.

[Reeuwijk .Reeuwijk .2010] reeuwijk2010osteoarthritisReeuwijk, KG., de Rooij, M., van Dijk, GM., Veenhof, C., Steultjens, MP. Dekker, J. 2010. Osteoarthritis of the hip or knee: which coexisting disorders are disabling? Osteoarthritis of the hip or knee: which coexisting disorders are disabling? Clinical rheumatology297739–747.

[Reich, Price PattersonReich .2008] reich2008principalReich, D., Price, AL. Patterson, N. 2008. Principal component analysis of genetic data Principal component analysis of genetic data. Nature genetics405491–492.

[RichardRichard2003] richard2003gibbsRichard, A. 2003. Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al Gibbs, john w belmont, paul hardenbol, thomas d willis, fuli yu, huanming yang, lan-yang ch'ang, wei huang, bin liu, yan shen, et al. The international hapmap project. Nature4266968789–796.

[Scheet StephensScheet Stephens2008] scheet2008linkageScheet, P. Stephens, M. 2008. Linkage disequilibrium-based quality control for large-scale genetic studies Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS genetics48.

[Schurz .Schurz .2019] schurz2019evaluatingSchurz, H., Müller, SJ., Van Helden, PD., Tromp, G., Hoal, EG., Kinnear, CJ. Möller, M. 2019. Evaluating the accuracy of imputation methods in a five-way admixed population Evaluating

the accuracy of imputation methods in a five-way admixed population. Frontiers in genetics1034.

[Shah .Shah .2020] shah2020genomeShah, S., Henry, A., Roselli, C., Lin, H., Sveinbjörnsson, G., Fatemifar, G.others 2020. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure Genome-wide association and mendelian randomisation analysis provide insights into the pathogenesis of heart failure. Nature Communications1111–12.

[Tachmazidou .Tachmazidou .2019] tachmazidou2019identificationTachmazidou, I., Hatzikotoulas, K., Southam, L., Esparza-Gordillo, J., Haberland, V., Zheng, J.others 2019. Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of uk biobank data. Nature genetics512230–236.

[Tam .Tam .2019] tam2019benefitsTam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. Meyre, D. 2019. Benefits and limitations of genome-wide association studies Benefits and limitations of genome-wide association studies. Nature Reviews Genetics208467–484.

[Tewhey, Bansal, Torkamani, Topol SchorkTewhey .2011] tewhey2011importanceTewhey, R., Bansal, V., Torkamani, A., Topol, EJ. Schork, NJ. 2011. The importance of phase information for human genomics The importance of phase information for human genomics. Nature Reviews Genetics123215–223.

[Thakker, Whyte, Eisman IgarashiThakker .2017] thakker2017geneticsThakker, RV., Whyte, MP., Eisman, J. Igarashi, T. 2017. Genetics of bone biology and skeletal disease Genetics of bone biology and skeletal disease. Academic Press.

[Tishkoff .Tishkoff .1996] tishkoff1996globalTishkoff, SA., Dietzsch, E., Speed, W., Pakstis, AJ., Kidd, JR., Cheung, K.others 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. Science27152541380–1387.

[WagnerWagner2013] wagner2013rareWagner, MJ. 2013. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. Pharmacogenomics144413–424.

[Wallace .Wallace .2017] wallace2017kneeWallace, IJ., Worthington, S., Felson, DT., Jurmain, RD., Wren, KT., Maijanen, H.Lieberman, DE. 2017. Knee

osteoarthritis has doubled in prevalence since the mid-20th century Knee osteoarthritis has doubled in prevalence since the mid-20th century. Proceedings of the National Academy of Sciences114359332–9336.

[Welter .Welter .2014] welter2014nhgriWelter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H.others 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic acids research42D1D1001–D1006.

[Zhang .Zhang .2008] zhang2008oarsiZhang, W., Moskowitz, R., Nuki, G., Abramson, S., Altman, RD., Arden, N.others 2008. OARSI recommendations for the management of hip and knee osteoarthritis, Part II: OARSI evidence-based, expert consensus guidelines Oarsi recommendations for the management of hip and knee osteoarthritis, part ii: Oarsi evidence-based, expert consensus guidelines. Osteoarthritis and cartilage162137–162.

[Zhu, Pan, Gu, Bradbury ChenZhu .2002] zhu2002aminoZhu, H., Pan, S., Gu, S., Bradbury, EM. Chen, X. 2002. Amino acid residue specific stable isotope labeling for quantitative proteomics Amino acid residue specific stable isotope labeling for quantitative proteomics. Rapid Communications in Mass Spectrometry16222115–2123.

Genetic Terminology

Human nucleus: Has 46 chromosomes (each double-strand DNA): 22 pairs of autosomes, and 2 sex chromosomes, X and/or Y.

Nuclear genome: DNA of these (22+X+Y) chromosomes, 3 109 bp.

Nuclear genome:

Locus: Position on a chromosome, or DNA at that position, or the piece of DNA coding for a trait.

Allele: Type of the DNA at a particular locus

SNP: Single nucleotide polymorphism; two alleles A and B.

Genotype: Pair of alleles (unordered) at a particular locus in a particular individual. AA, AB, BB

Homozygote: A genotype with two alike alleles. AA, BB

Heterozygote: A genotype with two different alleles. AB

Phenotype: Observable characteristics of an individual

Gene: The chunk of DNA coding for a functional protein.

Deletion: An abnormality in which part of a chromosome (carrying genetic material) is lost.

Diploid number of chromosomes: The number of chromosomes found in somatic cells, which in humans is 46.

DNA probe: A cloned DNA molecule labelled with a radioactive isotope (e.g., 32P or 35S) or a nonisotopic label (e.g., biotin). Used in molecular genetics to identify complementary DNA sequences by hybridizing to them.

DNA sequence: The relative order of base pairs, whether in a DNA fragment, gene, chromosome, or an entire genome.

Dominant: An allele that is almost always expressed, even if only one copy is present.

Exon: The protein-coding DNA sequence of a gene

Genetic polymorphism: Difference in DNA sequence among individuals, groups, or populations (e.g., genes for blue eyes versus brown eyes).

Genetic variation: A phenotypic variability of a trait in a population attributed to genetic heterogeneity.

Haploid number of chromosomes: The number of chromosomes found in sex cells, which in humans is 23.

Nuclear genome: The type of cell division that occurs in sex cells by which gametes having the haploid number of chromosomes are produced from diploid cells.

Monozygotic twins: Twins derived from a single fertilized ovum, i.e., identical twins.

Nucleic acids: Polymers of phosphorylated nucleosides, the building blocks of DNA and RNA.

Polymorphism: The existence of two or more different phenotypes resulting from two or more alleles, each with an appreciable frequency. Most blood group systems are polymorphic.

Population genetics: The branch of genetics that deals with how genes are distributed in populations and how gene and genotype frequencies stay constant or change.

Translation: The process of translating the codon sequence in mRNA into polypeptides with the help of tRNA and ribosomes.

X-chromosome: The sex chromosome present in double dose in females (XX) and in single dose in males (XY).

X-linked: Genes on the X chromosome, e.g., genes for hemophilia A, hemophilia B, and Xga blood group genes.

Y-chromosome: The sex chromosome present only in males (XY).

Appendix A

Appendix

As an additional document, you can find the code used to compile this thesis and the exact Bash and R commands. The same file has also been published in the following directory: `https://rpubs.com/fondan/quality_control_of_genotype_data`.

HelmholtzZentrum münchen
German Research Center for Environmental Health

UNIVERSITY OF THESSALY
creative years

Quality Control of genotype data from three different Illumina arrays:

Appendix A:

PLINK & R practical guideline

Xenofon Giannoulis, MSc.

Table of Contents

Introduction

Aims of the study

In this study, we have obtained peripheral whole blood sample to extract DNA for genotyping from 250 patients (104 males, 146 females) undergoing total joint replacement surgery from four different cohorts. One of the most computationally demanding challenges of this master thesis is merging cohorts from different array versions. Because of the complicated genetic assumptions of Illumina genotyping data, special handling of the initial files need to be made before proceeding to the Quality Control filters. In order to generate a high-quality Illumina genotyping merged data-set, we extract only the common variants across the different array types. We present the application of rigorous QC with PLINK where 9 individuals and 10,354 SNPs did not pass the QC criteria. The cleaned data set proceed for imputation using the Haplotype Reference Consortium in the Michigan server, incorporating the Minimac 4 algorithm for imputation and Eagle v2.4 for phasing. The final output consist of 39,127,678 SNPs without any additional filter applied.

Description of the Cohorts

FunGen CoreExome chip PLINK files, consisted by 3 different types of Exome Arrays :

Sample Type & Array Version	Samples	Build
HumanCoreExome-12v1-1_A	12	37
InfiniumCoreExome-24v1-1_A	34	37
InfiniumCoreExome-24v1-1_A	88	37
InfiniumCoreExome-24v1-3_A1	116	37

Total: 250 samples in 4 different datasets

Storage location:

```
/your_storage_location/user/your_name/4_QC
```

Plink Original File names

```
Cohort1.bed #14492
Cohort1.bim #542585 variants
Cohort1.fam #12 people

Cohort2.bed #81892
Cohort2.bim #551839 variants
Cohort2.fam #34 people

Cohort3.bed #109727
Cohort3.bim #551839 variants
Cohort3.fam #88people

Cohort4.bed #143648
Cohort4.bim #551004 variants
Cohort4.fam #116 people
```

Plink Version

```
PLINK v1.90b6.13 64-bit (30 Nov 2019)     <www.cog-genomics.org/plink/1.9/>
(C) 2005-2019 Shaun Purcell, Christopher Chang GNU General Public License v3
```

R version

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
```

Before Quality Control

Using only GenCall Genotypes

The GenCall application incorporates the GenTrain clustering and a calling a algorithm, where each SNP is analyzed independently to identify genotypes. The calling algorithm uses DNA intensity values and the output of the clustering algorithm to identify to which cluster the data for any specific locus corresponds.

The calling operation (classification) is performed using a Bayesian model and the GenCall score is the product of the GenTrain and a data-to-model fit score, where a score below 0.2 is considered as failed call rate and a score above 0.7 is reported as well behaving genotype.

Additional genotype calling methods are CRLMM, Illuminus and GenoSNP. There are factors such sample size or chromosome X inclusion in Quality Control that affect the decision of which method is the best.

Update maternal and paternal info from -9 to 0

The value of -9 is considered from PLINK software as "missing", so we recode the 3rd and the 4rth column of the fam file (which corresponds to paternal and maternal id respectively) to 0, which is consindered from PLINK as "not-provided" but does not causes any problems while running commands.

Saving the original files before we make any major change on the core file is common technique that it is used downstream multiple times.

Save original files as ORIGINAL

```
cp Cohort1.fam Cohort1.ORIGINAL.fam

cp Cohort2.fam Cohort2.ORIGINAL.fam

cp Cohort3.fam Cohort3.ORIGINAL.fam

cp Cohort4.fam Cohort4.ORIGINAL.fam
```

Run commands

I have samples with missing sex (again we recode the values from -9 to 0), usign awk and sponge for replacement:

```
cat Cohort1.fam | awk '{$3=0;$4=0;print}' | sponge Cohort1.fam

cat Cohort2.fam | awk '{$3=0;$4=0;print}' | sponge Cohort2.fam

cat Cohort3.fam | awk '{$3=0;$4=0;print}' | sponge Cohort3.fam

cat Cohort4.fam | awk '{$3=0;$4=0;print}' | sponge Cohort4.fam
```

Update 10 SNPs to match Fludigm annotation

Fluidigm provides sample identity and integrity checks of genomic data, by validating variant call reproducibility in the fluidigm and ABI 7900 system. SNP Genotyping concordance is defined as an allele call, that is identical to the previous determined allele type for the sample.(F.Yamatomo El at 2015)

Info

The following 10 SNPS have been changed to match Fluidigm id documentation. So I need to change these ids back to the original so that I can update using Wills Listing file. The following variants needs to be changed from rs to exm id in the cohort 2 and 3. The output file-names have "update10odds" to identify which step was run at what point.

```
Cohort2
Cohort3
```

This website provides more information: <heck this website for more info: http://grch37.ensembl.org/Homo_sapiens/Variation/Explore?r=12:109937034-109938034;v=rs7298565;vdb=variation;vf=395824877

2	exm193175	0	49189921	G	A
2	exm250297	0	182543455	A	G
2	exm261805	0	211060050	C	A
2	exm277163	0	234601669	C	A
5	exm466945	0	90151589	G	A
11	exm893239	0	17408630	G	A
12	exm1035579	0	109937534	G	A
15	exm1185460	0	86123988	G	A
15	exm1185487	0	86124555	G	A
17	exm1278107	0	3480447	G	A

I saved the ids in a file named "out_2columns.txt"

Run Commands

```
plink -bfile Cohort2  --update-name out_2_columns.txt --make-bed --out Cohort
2.update10odds

plink -bfile Cohort3  --update-name out_2_columns.txt --make-bed --out Cohort
2update10odds
```

Remove samples not part of the collection

These 30 samples needs to be excluded prior to QC, saved as "30_SampleExclusionsPriorQC.txt" :

- ARCO:

```
urn:wtsi:493153_D04_hip-fungen6636038   urn:wtsi:493153_D04_hip-fungen6636038
urn:wtsi:493153_E04_hip-fungen6636039   urn:wtsi:493153_E04_hip-fungen6636039
urn:wtsi:493153_F04_hip-fungen6636040   urn:wtsi:493153_F04_hip-fungen6636040
urn:wtsi:493153_G04_hip-fungen6636041   urn:wtsi:493153_G04_hip-fungen6636041
urn:wtsi:493153_H04_hip-fungen6636042   urn:wtsi:493153_H04_hip-fungen6636042
urn:wtsi:493153_A05_hip-fungen6636043   urn:wtsi:493153_A05_hip-fungen6636043
urn:wtsi:493153_B05_hip-fungen6636044   urn:wtsi:493153_B05_hip-fungen6636044
urn:wtsi:493153_C05_hip-fungen6636045   urn:wtsi:493153_C05_hip-fungen6636045
urn:wtsi:493153_D05_hip-fungen6636046   urn:wtsi:493153_D05_hip-fungen6636046
urn:wtsi:493153_E05_hip-fungen6636047   urn:wtsi:493153_E05_hip-fungen6636047
urn:wtsi:493153_F05_hip-fungen6636048   urn:wtsi:493153_F05_hip-fungen6636048
```

- 17 HIPS

  urn:wtsi:493153_D01_OAfg6273390 urn:wtsi:493153_D01_OAfg6273390
  urn:wtsi:493153_B01_OAfg6273380 urn:wtsi:493153_B01_OAfg6273380
  urn:wtsi:433501_A01_OAfg6273378 urn:wtsi:433501_A01_OAfg6273378
  urn:wtsi:493153_C01_OAfg6273388 urn:wtsi:493153_C01_OAfg6273388
  urn:wtsi:433501_G01_OAfg6273384 urn:wtsi:433501_G01_OAfg6273384
  urn:wtsi:493153_A01_OAfg6273378 urn:wtsi:493153_A01_OAfg6273378
  urn:wtsi:433501_C01_OAfg6273380 urn:wtsi:433501_C01_OAfg6273380
  urn:wtsi:433501_C02_OAfg6273388 urn:wtsi:433501_C02_OAfg6273388
  urn:wtsi:433501_E02_OAfg6273390 urn:wtsi:433501_E02_OAfg6273390
  urn:wtsi:433501_D01_OAfg6273381 urn:wtsi:433501_D01_OAfg6273381
  urn:wtsi:433501_E01_OAfg6273382 urn:wtsi:433501_E01_OAfg6273382
  urn:wtsi:433501_F01_OAfg6273383 urn:wtsi:433501_F01_OAfg6273383
  urn:wtsi:433501_H01_OAfg6273385 urn:wtsi:433501_H01_OAfg6273385
  urn:wtsi:433501_A02_OAfg6273386 urn:wtsi:433501_A02_OAfg6273386
  urn:wtsi:433501_B02_OAfg6273387 urn:wtsi:433501_B02_OAfg6273387
  urn:wtsi:433501_D02_OAfg6273389 urn:wtsi:433501_D02_OAfg6273389
  urn:wtsi:433501_F02_OAfg6273391 urn:wtsi:433501_F02_OAfg6273391

- 1 Unknown

  urn:wtsi:433501_B01_OAfg6273379 urn:wtsi:433501_B01_OAfg6273379

- 1 Self-Annoted Asian (info from the manifest)

  urn:wtsi:433501_E03_OAfg6308472 urn:wtsi:433501_E03_OAfg6308472

In Plink, for any data manipulation, it is important to save files on tab delimited format. The sample ids need always to be saved on the first and second column of the file.

- Commands

```
plink --bfile Cohort2.update10odds --remove 30_SampleExclusionsPriorQC.tx
t --make-bed --out Cohort2_update10odds.30samplesout
```

```
plink --bfile Cohort3.update10odds --remove 30_SampleExclusionsPriorQC.tx
t --make-bed --out Cohort3.update10odds.30samplesout
```

Update position and strand using strand files

Since I have 3 different array types, before merging my datasets I need to update Strand & Position on each file according to the version of the corresponding array. Files can be found here: https://www.well.ox.ac.uk/~wrayner/strand/index.html

*Commands:

```
./update_build.sh Cohort1 your_storage_location/your_name/xenophon/_____5_St
rand_files_____/humancoreexome-12v1-1_a-b37.strand Cohort1.strandupdated

./update_build.sh Cohort2.update10odds.30samplesout your_storage_location/use
r/your_name/_____5_Strand_files_____/InfiniumCoreExome-24v1-1_A-b37.strand
Cohort2.update10odds.13samplesout.strandupdated

./update_build.sh Cohort3.update10odds.30samplesout your_storage_location/use
r/your_name/_____5_Strand_files_____/InfiniumCoreExome-24v1-1_A-b37.strand
Cohort3.update10odds.13samplesout.strandupdated

./update_build.sh Cohort4 your_storage_location/user/your_name/_____5_Strand
_files_____/InfiniumCoreExome-24v1-3_A1-b37.strand Cohort4.strandupdated
```

Update Pseudoautosomal (PAR1 & PAR2) regions

Pseudoautosomal Regions (PAR1 and PAR2) of the X and Y Chromosomes pair and recombine during meiosis like autosomes, but the recombination activity in PAR 1 is different between sexes (A. Flaquer El at 2008). PAR 1 is located at the terminal region of the short arms and spans the first 2.7Mb (rough estimate) of the proximal arm of the human sex Chromosomes. On the other hand, PAR2 encompasses the distal 320kb of the long arm of each sex chromosomes (D J. Cotter 2016), resulted from at least two duplications from the X chromosome to the terminal end of the Y chromosome(Charchar et al. 2003).

To date, at least 24 expressed genes have be identified in the PAR1 region (A Helena Mangs and Brian J Morris, 2007) and at least 5 in PAR2 region (Rüdiger Jörg, Blaschke GudrunRappold, 2016).

Autosome codes in PLINK are considered from 1 through 22. X chromosome's pseudo-autosomal region is annotated as a seperate 'XY Chromosome' which has number code 25.

```
X    X chromosome                  -> 23
Y    Y chromosome                  -> 24
XY   Pseudo-autosomal region of X  -> 25
MT   Mitochondrial                 -> 26
```

In this step, we update the pseudoautosomal region by removing the need of special handling of male heterozygous calls. For each dataset we update the PAR1/2 and the the X and Y Chromosome to chr25. –split-x takes the base pair position boundaries of the pseudo-autosomal region, and changes the chromosome codes of all variants in the region to XY.

Note: I export the files without the 10 odds and 30 samples included in the file-names in order to be consistent with the workflow

Run commands:

```
plink --bfile Cohort1.strandupdated --split-x b37 --make-bed --out Cohort1.st
randupdated.PARupdated

plink --bfile Cohort2.update10odds.30samplesout.strandupdated --split-x b37 -
-make-bed --out Cohort2.strandupdated.PARupdated

plink --bfile Cohort3.update10odds.30samplesout.strandupdated --split-x b37 -
-make-bed --out Cohort3.strandupdated.PARupdated

plink --bfile Cohort4.strandupdated --split-x b37 --make-bed --out Cohort4.st
randupdated.PARupdated
```

Problems and Solutions before merging

Problem Information

Before start any data manipulation, we create two files. The first one is called "IDListing.txt" having 590,623 official Illumina ids from all arrays involved and how many times each id was present for each array respectively. The second file is called "ProbeListing.txt" containing 553,860 probe sequences from all arrays involved and how many times each probe sequence is present across the 3 different arrays. The idea is to compare those files and identify possible mismatches in terms of variants' counts and identification.

`/your_storage_location/user/your_name/1_1_4_Tec`

- IDListing.txt contains the probe id sequence of each array and the amount of times a probe sequence was counted across the 3 different array versions. Looks like this:

Variant id & chr:pos	Count	-12v1-1_A	-24v1-1_A	-24v1-3_A1
rs2566-131_T_R_1891342930	3	rs2566	rs2566	rs2566

- ProbeListing.txt contains the probe sequence of each array and the amount of times a probe sequence was counted across the 3 different arrays version. Looks like this:

Probe	Count	-12v1-1_A	-24v1-1_A	-24v1-3_A1
GTCAACAGCAGAGTGTGTATAGCTG GTCAACAAAACGCTAAAACCACAGG	3	exm828933	exm828933	exm828933

When we listed per chr:pos we found:

- Problem 1: We noticed that there were cases where even if the count of the second column was 3, the rs_id name was missing from the actual array list (probably duplicates on one array list):

Probe	Count	-12v1-1_A	-24v1-1_A	-24v1-3_A1
TTGAGCACTCTTCTGTAATCTCATGA GGTGTCCAGGGAAGAGACAATGAT	3	rs1873983	rs1873983	?
TTTCTTTCGCATTGCAATGCCATGCT CCCTGCTCTTGGTCCTGTTTTCTA	3	exm1261061	rs1469122	?

- Problem 2: We found same the rs_ids between FunGen 2017 and 2018, but different chromosome position - 1 base pair difference:

rs_id	Morgans or Centimorgans	Base Pair	Allele A	Allele B	FunGen File
indel.20904	0	102650052	I	D	coreex_OA18DNA_20180726
indel.20904	0	102650053	I	D	coreex_fungen_20170807

Which was included because in the ProbeListing.txt and IDListing file it was displayed like this:

- ProbeListing:

Probe	Count	-12v1-1_A	-24v1-1_A	-24v1-3_A1
TCTCCTCCAGGATTGTGAATTATAC ACCAGATTTGCCAAGAGATGCTGTT	3	indel.2090	indel.2090	indel.2090

- IDListing:

rs_id	Count	Base Pair	-12v1-1_A	-24v1-1_A	-24v1-3_A1
indel.20904_IlmnDup-1_P_R_2358125101	1	102650052	?	?	indel.20904
indel.20904-1_M_R_2113295257	2	102650053	indel.20904	indel.20904	?

- Problem 3: We found in the 2018 FunGen file variants with different rs_ids but same chromosome position.

rs_id	Morgans or Centimorgans	Base Pair	Allele A	Allele B
exm2253593	0	900427	A	G
exm596	0	900427	A	G

- Problem 4: On the same chr:position we found different rs_ids, different sequence probes and multiallelic variants

Solving Information

- Create the Inclusion List 'Finallistwith3.txt' The best way to make sure that any variant is present across the 3 arrays was to extract from the probe listing file only the 3s with all 3 fields not null

```
cat /your_storage_location/user/your_name/1_1_4_Tec/ProbeListing.txt | awk '$2==3 && !($3!="" || $3!="" || $4=="")' > /your_storage_location/user/your_name/4_QC
```

- Join IDs and positions in all 4 bim files and save them as "merged.bims.txt"

```
join -1 1 -2 2 <(join -1 1 -2 2 <(join -j2 <(sort -k2,2 Cohort1.strandupdated.PARupdated.bim) <(sort -k2,2 Cohort2.strandupdated.PARupdated.bim)) <(sort -k2,2 Cohort3.strandupdated.PARupdated.bim)) <(sort -k2,2 Cohort4.strandupdated.PARupdated.bim) > merged.bims.txt
```

- Create the "inclusion.lists.txt" by removing from the join above list the ones that have a different position in any of the 4 files

```
fgrep -w -v -f <(awk '!($4==$9 && $9==$14 && $14==$19)' merged.bims.txt | cut -d' ' -f1) /your_storage_location/user/your_name/4_QC/Finallistwith3.txt > inclusion.lists.txt
```

Extract the "Survived" Variants from each dataset

Here we extract the inclusion.list.txt from each funGen file in order to merge in the next step only the common variants across the 3 array types.

```
--bfile Cohort1.strandupdated.PARupdated --extract inclusion.lists.txt --make
-bed --out Cohort1.strandupdated.PARupdated.Extracted

--bfile Cohort2.strandupdated.PARupdated --extract inclusion.lists.txt --make
-bed --out Cohort2.strandupdated.PARupdated.Extracted

--bfile Cohort3.strandupdated.PARupdated --extract inclusion.lists.txt --make
-bed --out Cohort3.strandupdated.PARupdated.Extracted

--bfile Cohort4.strandupdated.PARupdated --extract inclusion.lists.txt --make
-bed --out Cohort4.strandupdated.PARupdated.Extracted
```

Recode rs_ids to chromosome and position

From:

1	exm2216284	0	564862	T	C
1	MitoG4821A	0	565370	A	G
1	MitoA5657G	0	566206	G	A
1	MitoA5952G	0	566501	G	A

To:

1	1:564862_T_C	0	564862	T	C
1	1:565370_A_G	0	565370	A	G
1	1:566206_G_A	0	566206	G	A
1	1:566501_G_A	0	566501	G	A

We will save original files as ORIGINAL and I will use AWK command to recode the bim files from rs_ids to chromosome position

```
cp Cohort1.strandupdated.PARupdated.Extracted.bim Cohort1.strandupdated.PARup
dated.Extracted.Original.bim

cp Cohort2.strandupdated.PARupdated.bim Cohort2.strandupdated.PARupdated.Orig
inal.bim

cp Cohort3.strandupdated.PARupdated.bim Cohort3.strandupdated.PARupdated..Ori
ginal.bim

cp Cohort4.strandupdated.PARupdated.bim Cohort4.strandupdated.PARupdated.Orig
inal.bim
```

- I use awk to transform the bim files per chromosome position:

```
    awk 'BEGIN{FS="\t"; OFS="\t"}{if($5<$6){first=$5;second=$6;}else{first=$6
;second=$5;} print $1,$1":"$4"_"first"_"second, $3, $4,$5,$6}' Cohort1.strand
updated.PARupdated.Extracted.Original.bim > Cohort1.strandupdated.PARupdated.
Extracted.bim

    awk 'BEGIN{FS="\t"; OFS="\t"}{if($5<$6){first=$5;second=$6;}else{first=$6
;second=$5;} print $1,$1":"$4"_"first"_"second, $3, $4,$5,$6}' cCohort2.stran
dupdated.PARupdated.ORIGINAL.bim > Cohort2.strandupdated.PARupdated.bim

    awk 'BEGIN{FS="\t"; OFS="\t"}{if($5<$6){first=$5;second=$6;}else{first=$6
;second=$5;} print $1,$1":"$4"_"first"_"second, $3, $4,$5,$6}' Cohort3.strand
updated.PARupdated.ORIGINAL.bim > Cohort3.strandupdated.PARupdated.bim

    awk 'BEGIN{FS="\t"; OFS="\t"}{if($5<$6){first=$5;second=$6;}else{first=$6
;second=$5;} print $1,$1":"$4"_"first"_"second, $3, $4,$5,$6}' Cohort4.strand
updated.PARupdated.ORIGINAL.bim > Cohort4.strandupdated.PARupdated.bim
```

- We identify and save all multiallelic and problematic variants (342) in the "multiallelic.exclusion.tailed.txt" which looks like this:

head multiallelic.exclusion.tailed.txt

1:2435759_G_T
1:2435759_G_A
1:15900221_D_I
1:15900221_A_T
1:16535435_C_T

tail multiallelic.exclusion.tailed.txt

22:26886166_C_T
23:118227947_I_D
23:118227947_C_T
23:134421551_I_D
23:134421551_G_C

The list was generated by making merge-testing and exporting the warnings with the following command:

```
    fgrep "have the same" merge.test.log | cut -d' ' -f3,5 |sed "s/'//g;s/ /\
n/" > multiallelic.exclusion.txt
```

- Extract variants that survive

```
    plink --bfile Cohort1.strandupdated.PARupdated --extract <(awk '{print $3
}' inclusion.lists.txt) --make-bed --out Cohort1.strandupdated.PARupdated.Ext
racted

    plink --bfile Cohort2.strandupdated.PARupdated --extract <(awk '{print $4
}' inclusion.lists.txt) --make-bed --out Cohort2.strandupdated.PARupdated.Ext
racted

    plink --bfile Cohort3.strandupdated.PARupdated --extract <(awk '{print $4
}' inclusion.lists.txt) --make-bed --out Cohort3.strandupdated.PARupdated.Ext
racted

    plink --bfile Cohort4.strandupdated.PARupdated --extract <(awk '{print $5
}' inclusion.lists.txt) --make-bed --out Cohort4.strandupdated.PARupdated.Ext
racted
```

- I save the ORIGINAL bim files

  cp Cohort1.strandupdated.PARupdated.Extracted.bim
  Cohort1.strandupdated.PARupdated.Extracted.ORIGINAL.bim

Merge Fungen files

- I create the merge_list.txt

```
Cohort1.strandupdated.PARupdated.Extracted

Cohort2.strandupdated.PARupdated.Extracted

Cohort3.strandupdated.PARupdated.Extracted

Cohort4.strandupdated.PARupdated.Extracted
```

- Merging Cohorts

```
--plink --merge-list merge_list.txt --make-bed --out Firstmerge_all_Cohorts
```

Creating the SNP List Directory

```
mkdir SNPlists
cd SNPlists

#!/bin/sh

for chr in $(seq 1 26) ; do  plink --bfile /your_storage_location/user/your_n
ame/4_QC/ Firstmerge_all_Cohorts --chr $chr --write-snplist --out chr$chr ; d
one
cat chr1.snplist chr2.snplist chr3.snplist chr4.snplist chr5.snplist chr6.snp
list chr7.snplist chr8.snplist chr9.snplist chr10.snplist chr11.snplist chr12
.snplist chr13.snplist chr14.snplist chr15.snplist chr16.snplist chr17.snplis
t chr18.snplist chr19.snplist chr20.snplist chr21.snplist chr22.snplist > chr
1_22.snplist
wc -l *snplist > wc.snplist
```

Sample QC Steps taken and commands in R

Edit and the 'Produce Plink Summary Files' script to match correspinding files and directories:

```
/your_storage_location/user/your_name/4_QC/ Firstmerge_all_Cohorts

# ./ProducePlinkSummaryFilesEditfor.sh /your_storage_location/user/your_name/
4_QC/Firstmerge_all_Cohorts


# cat ProducePlinkSummaryFiles

#!/bin/sh

# ProducePlinkSummaryFiles.sh
# A script to run through the preliminary PLINK steps for Core Exome Chip

# Required parameters:
# Directory
# Cohort name
# PLINK File name without the extension

# Usage is ./ProducePlinkSummaryFiles.sh <directory name> <cohort name> <Genc
all file name>

DIR=$1
COHORT=$2
FILE=$3

echo File used here is $DIR/$FILE

#Prefilter exclude samples and SNPs with  call rate >10% missing

plink --bfile $DIR/$FILE --allow-no-sex --geno 0.1 --mind 0.1 --make-bed --ou
t $DIR/$COHORT\_prefiltered

#Rename the file name
```

```
FILE="prefiltered"


## Call rate ##

# Run missingness across file genome-wide
plink --bfile $DIR/$COHORT\_$FILE --allow-no-sex --missing --out $DIR/$COHORT
\_$FILE-missing

# Produce a log file giving number of samples excluded at CR 0.98 to check ag
ainst R result
plink --bfile $DIR/$COHORT\_$FILE --allow-no-sex --mind 0.02 --make-bed --out
$DIR/$COHORT\_$FILE-mind0.02
rm $DIR/$COHORT\_$FILE-mind0.02.bed $DIR/$COHORT\_$FILE-mind0.02.bim $DIR/$CO
HORT\_$FILE-mind0.02.fam

## Sex check ##

# Run sex checking
plink --bfile $DIR/$COHORT\_$FILE --allow-no-sex --check-sex --out $DIR/$COHO
RT\_$FILE-sexcheck

# Extract xchr SNPs
plink --bfile $DIR/$COHORT\_$FILE --allow-no-sex --chr 23 --make-bed --out $D
IR/$COHORT\_$FILE-xchr

# Run missingness on xchr SNPs
plink --bfile $DIR/$COHORT\_$FILE-xchr --allow-no-sex --missing --out $DIR/$C
OHORT\_$FILE-xchr-missing

## Heterozygosity ##

# Extract autosomal SNPs - ! YOU NEED TO SUBSTITUTE YOUR OWN SNPLIST HERE!

plink --bfile $DIR/$COHORT\_$FILE --allow-no-sex --extract /your_storage_loca
tion/user/your_name/4_QC/ Firstmerge_all_Cohorts/SNPlists/chr1_22.snplist --m
ake-bed --out $DIR/$COHORT\_$FILE-chr1-22

# Extract SNPs with MAF ≥1%

plink --bfile $DIR/$COHORT\_$FILE-chr1-22 --allow-no-sex --maf 0.01 --make-be
d --out $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01
```

```
# Extract SNPs with MAF <1%

plink --bfile $DIR/$COHORT\_$FILE-chr1-22 --allow-no-sex --exclude $DIR/$COHO
RT\_$FILE-chr1-22-mafgte0.01.bim --make-bed --out $DIR/$COHORT\_$FILE-chr1-22
-mafless0.01


# Get missingness to plot against het in R

plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafless0.01 --allow-no-sex --missin
g --out $DIR/$COHORT\_$FILE-chr1-22-mafless0.01-missing

plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01 --allow-no-sex --missing
--out $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-missing

# Convert both to ped/map files for Wills het script

plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafless0.01 --allow-no-sex --recode
--out $DIR/$COHORT\_$FILE-chr1-22-mafless0.01-recode

plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01 --allow-no-sex --recode
--out $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-recode
```

YOU NEED TO SUBSTITUTE YOUR OWN DIRECTORY

```
# I use the following for running het script:

echo "perl /your_storage_location/user/your_name/4_QC/Scripts/calc_het.pl -f
$DIR/$COHORT\_$FILE-chr1-22-mafless0.01-recode.ped"

echo "perl /your_storage_location/user/your_name/4_QC/Scripts/calc_het.pl -f
$DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-recode.ped"

## Relatedness/Duplicates ##
# Pair-wise IBD to look at duplicates.
# Using only autosomal variants ≥1%, excluding complex regions and LD prune u
sing R-squared 0.2.
# Exclude complex regions (these are regions chr and position so this file ca
n be used with any b37 array data)
```

```
plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01 --allow-no-sex --exclude
range /your_storage_location/user/your_name/4_QC/ Firstmerge_all_Cohorts/Scri
pts/complex_regions.txt --make-bed --out $DIR/$COHORT\_$FILE-chr1-22-mafgte0.
01-noCR

#LD prune
plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-noCR --allow-no-sex --in
dep 50 5 1.25 --out $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-noCR-pruning

#Count SNPs in each file for the log
wc -l *prune.*

#Extract only the prune in SNPs
plink --bfile $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-noCR --allow-no-sex --ex
tract $DIR/$COHORT\_$FILE-chr1-22-mafgte0.01-noCR-pruning.prune.in --make-bed
--out $COHORT\_$FILE-chr1-22-mafgte0.01-noCR-LDpruned0.2

#Run genome
plink --bfile $COHORT\_$FILE-chr1-22-mafgte0.01-noCR-LDpruned0.2 --genome --o
ut $COHORT\_$FILE-chr1-22-mafgte0.01-noCR-LDpruned0.2.genome
```

- Run the script in the UNQCed merged files to apply the below mentioned filters:

```
./ProducePlinkSummaryFiles.sh /your_storage_location/user/your_name/4_QC/Firs
tmerge_all_Cohorts
```

Run the space to tab script

Here I run the space_to_tab.pl script on the individual missingness and locus missingess files as well as the rest files I am going to use downstream, especially when using R as visualization tool.

```
perl space_to_tab.pl Firstmerge_all_Cohorts-missing.lmiss
perl space_to_tab.pl Firstmerge_all_Cohorts-missing.imiss
perl space_to_tab.pl Firstmerge_all_Cohorts_prefiltered-missing.lmiss
perl space_to_tab.pl Firstmerge_all_Cohorts_prefiltered-missing.imiss
```

```
perl space_to_tab.pl Firstmerge_all_Cohorts-sexcheck.sexcheck
perl space_to_tab.pl Firstmerge_all_Cohorts_prefiltered-sexcheck.sexcheck
perl space_to_tab.pl Firstmerge_all_Cohorts_prefiltered-xchr-missing.imiss
```

I create the following files which I load on my jupyter R notebook located in this directory:

- tab- `Firstmerge_all_Cohorts` -missing.lmiss

- tab- `Firstmerge_all_Cohorts` -missing.imiss

R COMMANDS FOR VISUALIZATION ANDINTERPRETATION OF THE RESULTS

Use Gencall Only and load individual missingness file

```
imiss <- read.table('…/tab- Firstmerge_all_Cohorts-missing.imiss', header = T
RUE, stringsAsFactors = FALSE)
```

Apply a generalized threshold to exclude the bad quality samples/individual with more than 10 % missingness

```
x<- which(imiss$F_MISS > 0.1)
length(x)

#Make a histogram

hist(imiss$F_MISS, freq=TRUE, col="light blue", border= "dark green", main =
"Sample Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS", ylab = "
Frequency", ylim=c(0,100))

#Zoom in on the histogram

hist(imiss$F_MISS, freq=TRUE, col="light blue", border= "dark green", main =
"Sample Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS", ylab = "
Frequency", ylim = c(0,10))
abline(v=0.02, col="dark red",lwd = 2, lty = 2)

y <- which(imiss$F_MISS >0.09)
y
```

Sample Call rate < 98%

Individual Call/Missingness Rate –missing

Missingness or Call Rate is the proportion of missing genotypes after algorithm calling per individual. It is an indicator of SNP quality from the original genotyping array and usually plotted along with heterozygosity rate because they affect each other and define the quality of the genotyped sample.

A recommended threshold for removing individuals with at least 98% call rate.

```
imiss_call_rate <- read.table('…/tab-Firstmerge_all_Cohorts -missing.imiss',
header = TRUE, stringsAsFactors = FALSE)

summary (imiss_call_rate$F_MISS)

hist(imiss_call_rate$F_MISS, freq=TRUE, col="lightblue3", border= "dark green
", main = "Sample Call Rate", sub = " Firstmerge_all_Cohorts", xlab = "F_MISS
", ylab = "Frequency", ylim=c(0,90))

#zoom in to highlight the tails of the histogram:

hist(imiss_call_rate$F_MISS, freq=TRUE, col="lightblue3", border= "dark green
", main = "Sample Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS"
, ylab = "Frequency", ylim=c(0,10))
abline(v=0.02, col="dark red",lwd = 2, lty = 2)

cr1<-which(imiss_call_rate$F_MISS > 0.02 )#98% call rate
length(cr1)

# 0 samples to exclude but we will save the output in case there was an exclu
sion:

CR1<- imiss_call_rate [cr1,]
crnumber1<- dim(CR1)
TEMP1<- imiss[cr1,2]
length(cr1)
```

```
# Save failed called rates:

write.table(CR1, "callrate_fails.txt", sep="\t", col.names=FALSE, row.names=F
ALSE, quote=FALSE)
```

Sex checking

- Useful to see if there has been a plate rotation in the genotyping
- Heterozygosity check of the X Chr
- F statistic on the x-axis is the probability an individual inherited two identical alleles from a single ancestor, males should be close to 1, females close to 0.
- Color represents the self-reported gender of the participants and the actual points the gender indentified from the DNA genotyping

```
sexcheck <- read.table('…/tab-Firstmerge_all_Cohorts_prefiltered-sexcheck.sex
check', header = TRUE, stringsAsFactors = FALSE)

names(sexcheck)  # [1] "FID"    "IID"    "PEDSEX" "SNPSEX" "STATUS" "F"
dim(sexcheck)
str(sexcheck)
head(sexcheck)

xchr_imiss <- read.table('…/tab-Firstmerge_all_Cohorts_prefiltered-xchr-missi
ng.imiss', header = TRUE, stringsAsFactors = FALSE)

sexcheck_imiss <- data.frame(FID=sexcheck$FID,IID=sexcheck$IID, PEDSEX=sexche
ck$PEDSEX, SNPSEX=sexcheck$SNPSEX, STATUS=sexcheck$STATUS, F_inbreed=sexcheck
$F, F_MISS=xchr_imiss$F_MISS)
```

```
#PLOTING TO MAKE A DECISION:


plot(sexcheck_imiss$F_inbreed, xchr_imiss$F_MISS, col="grey", main="Sex check
", sub= "Fungen1-Fungen2-QQ-QQ2", xlab="X chr inbreeding (homozygosity) estim
ate F", ylab="Proportion of missing SNPs for the X chr")
temp <- subset(sexcheck_imiss, sexcheck_imiss$PEDSEX=="1") #1=males
points(temp$F_inbreed, temp$F_MISS, col="Lightgreen")
temp <- subset(sexcheck_imiss, sexcheck_imiss$PEDSEX=="2") #2=females
points(temp$F_inbreed, temp$F_MISS, col="dark red")
temp <- subset(sexcheck_imiss, sexcheck_imiss$STATUS=="PROBLEM") #STATUS
points(temp$F_inbreed, temp$F_MISS, col="Yellow", pch=16,cex=0.8)
legend("topright", c("Male PEDSEX","Female PEDSEX", "Problem Status"), fill=c
("Lightgreen","dark red", "Yellow"))
identify(sexcheck_imiss$F_inbreed, sexcheck_imiss$F_MISS, labels=sexcheck_imi
ss$FID)

#PLOTING SAME RESULTS WITH DIFFERENT GRAPH

# Define color and point types depending on Gender
pch_vec <- rep(19, times = nrow(sexcheck_imiss))
pch_vec[sexcheck_imiss$PEDSEX=="1"] <- 17

col_vec <- rep(rgb(1, 0, 0, alpha = 0.5), times = nrow(sexcheck_imiss))
col_vec[sexcheck_imiss$PEDSEX=="2"] <- rgb(0, 0, 1, alpha = 0.3)

# Start with empty plot
par(mar = c(5.1,6.1,2.1,7.1))


plot(sexcheck_imiss$F_inbreed,  xchr_imiss$F_MISS, pch = pch_vec, col = col_v
ec, type = "n",
     ylab="Proportion of missing SNPs for the X chr",
     xlab="X chr inbreeding (homozygosity) estimate F")

# Add points
points(sexcheck_imiss$F_inbreed, xchr_imiss$F_MISS, pch = pch_vec, col = col_
vec)

abline(lm(sexcheck_imiss$F_inbreed ~ xchr_imiss$F_MISS), col="blue")
abline(lm(sexcheck_imiss$F_inbreed ~ sexcheck_imiss$PEDSEX=="2"),col="red")
```

```
# Add legend
legend("right", legend = c("males", "females"),
       col = c(2,4), pch = c(17,19), lwd =2, lty =1,
       xpd = TRUE, inset = -0.30, title = "Gender")
```

Heterozygosity by MAF bin

Heterozygosity –het

Heterozygosity is fundamental to the study of genetic variation in populations. It is the degree of having two different alleles at a locus. Excess heterozygosity is possibly caused due to sample contamination and when identified less than expected heterozygosity rates sample inbreeding is the most common causal reason. The threshold is usually applied by removing individuals with 3 standard deviations from heterozygosity mean rate of all samples. It is common that genotyping individual call rate and heterozygosity call rate to be plotted together and cutoffs are selected so as to identify outlier individuals based on both statistical indicators.

Minor allele Frequency (MAF) –maf

MAF refers to the frequency at which the second most common allele occurs in a given population. This measure provides information that distinguish common polymorphisms (MAF>1%) from rare variants (MAF<1%). In other words, if there are 3 alleles, with frequencies of 0.50, 0.36 and 0.01, the MAF will be reported as 0.36. The database of Single Nucleotide Polymorphisms (dbSNP) https://www.ncbi.nlm.nih.gov/snp/ contains MAF information based on the 1000 Genomes Project Consortium.

During the analysis MAF is used to identify if self-reported gender matches are accurate based on the heterozygosity rates.

- MAF≥1%

```
dir <- "/…/"
Cohort <- "Firstmerge_all_Cohorts"
file <- "prefiltered-chr1-22-"
maf <- "mafgte0.01"

Het <-read.table(paste(dir, "Summary-", Cohort, "_", file, maf, "-recode.ped"
, sep=""), header=TRUE, fill=TRUE)
Aut_imiss<-read.table(paste(dir, Cohort, "_", file, maf, "-missing.imiss", se
p=""), header=TRUE, fill=TRUE) #autosomal call rate/missingness results

hist(Het$Percent_het, freq=TRUE, col="lightblue3", border= "dark green", main
= "Autosomal heterozygosity MAF≥1%", sub = paste(Cohort, maf, sep=" "), xlab=
"% Heterozygosity", ylab="Frequency")

#change y axis to focus in on tails

hist(Het$Percent_het, freq=TRUE, col="lightblue3", border= "dark green", main
= "Autosomal heterozygosity MAF≥1%", sub = paste(Cohort, maf, sep=" "), xlab=
"% Heterozygosity", ylab="Frequency", ylim= c(0,10))

#plotting het and call rate - the files are in the same order so can just com
bine columns.

Het_Aut_imiss <- data.frame(ID=Het$ID, percent_het=Het$Percent_het, ID2=Aut_i
miss$IID, F_MISS=Aut_imiss$F_MISS)

plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, main = "Autosomal heter
ozygosity and call rate", sub = paste(Cohort, maf, sep=" "),xlab="% Heterozyg
osity", ylab="F_MISS", ylim= c(0,0), xlim = c(34,39))
```

Find which individuals are 3 standard deviations away from the mean:

```
line1<-mean(Het$Percent_het, na.rm=T)-3*sd(Het$Percent_het, na.rm=T)
line2<-mean(Het$Percent_het, na.rm=T)+3*sd(Het$Percent_het, na.rm=T)

print(line1)
print(line2)

which(Het$Percent_het>line2)

Het$Percent_het[which(Het$Percent_het>line2)]
```

```
more_line2 <- which(Het$Percent_het>line2)
length(more_line2)
Het[more_line2,]



print(line1)
which(Het$Percent_het<line1)
Het$Percent_het[which(Het$Percent_het<line1)]
less_line1 <- which(Het$Percent_het<line1)
length(less_line1)
Het[less_line1,]

Exclu <- Het_Aut_imiss[c(more_line2, less_line1),]
dim(Exclu)
Exclu

plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, main = "Autosomal heter
ozygosity and call rate", sub = paste(Cohort, maf, sep=" ") ,xlab="% Heterozy
gosity", ylab="F_MISS", col = rgb(0.2, 0.6, 0.6, alpha = 0.7))
temp <- subset(Exclu)
points(temp$percent_het, temp$F_MISS, col="red", pch=16, cex=0.8)
legend("topright", c(">3xSD"), fill=c("red"))
identify(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, labels=Het_Aut_imis
s$ID)
```

Producing the same graph with different range:

```
plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, main = "Autosomal heter
ozygosity and call rate", sub = paste(Cohort, maf, sep=" ") ,xlab="% Heterozy
gosity", ylab="F_MISS", col = rgb(0.2, 0.6, 0.6, alpha = 0.7),  xlim = c(32,4
0))
temp <- subset(Exclu)
points(temp$percent_het, temp$F_MISS, col="red", pch=16, cex=0.8)
legend("topright", c(">3xSD"), fill=c("red"))
identify(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, labels=Het_Aut_imis
s$ID)
```

- MAF<1%

```
dir <- "/…/"
Cohort <- "Firstmerge_all_Cohorts"
file <- "prefiltered-chr1-22-"
maf <- "mafless0.01"

Het <- read.table(paste(dir, "Summary-", Cohort, "_", file, maf, "-recode.ped
", sep=""), header=TRUE, fill=TRUE)

Aut_imiss<-read.table(paste(dir, Cohort, "_", file, maf, "-missing.imiss", se
p=""), header=TRUE, fill=TRUE) #autosomal call rate/missingness results

names(Het)
names(Aut_imiss)

dim(Het)
dim(Aut_imiss)

hist(Het$Percent_het, freq=TRUE, col="green", border ="red", main = "Autosoma
l heterozygosity MAF<1%", sub = paste(Cohort, maf, sep=" "), xlab="% Heterozy
gosity", ylab="Frequency")

#change y axis to focus in on tails
hist(Het$Percent_het, freq=TRUE, col="green", border ="red", main = "Autosoma
l heterozygosity MAF <1%", sub = paste(Cohort, maf, sep=" "), xlab="% Heteroz
ygosity", ylab="Frequency", ylim= c(0,20))

# plotting het and call rate - the files are in the same order so can just co
mbine columns.
Het_Aut_imiss <- data.frame(ID=Het$ID, percent_het=Het$Percent_het, ID2=Aut_i
miss$IID, F_MISS=Aut_imiss$F_MISS)

dim(Het_Aut_imiss)

names(Het_Aut_imiss)

head(Het_Aut_imiss)
```

```
plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, col="dark blue", main =
"Autosomal heterozygosity and call rate", sub = paste(Cohort, maf, sep=" "),x
lab="% Heterozygosity", ylab="F_MISS", pch = 21, bg = 2)
identify(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, labels=Het_Aut_imis
s$ID)

#Checking to see if any samples are more than 3xSD away from the mean:

summary(Het$Percent_het)

mean(Het$Percent_het, na.rm=T)-3*sd(Het$Percent_het, na.rm=T)
mean(Het$Percent_het, na.rm=T)+3*sd(Het$Percent_het, na.rm=T)

line1<-mean(Het$Percent_het, na.rm=T)-3*sd(Het$Percent_het, na.rm=T)
line2<-mean(Het$Percent_het, na.rm=T)+3*sd(Het$Percent_het, na.rm=T)

print(line2)
which(Het$Percent_het>line2)

Het$Percent_het[which(Het$Percent_het>line2)]

more_line2 <- which(Het$Percent_het>line2)
length(more_line2)
Het[more_line2,]

print(line1) #mafgte0.01
which(Het$Percent_het<line1)

Het$Percent_het[which(Het$Percent_het<line1)]

less_line1 <- which(Het$Percent_het<line1)
length(less_line1)  #mafgte0.01 6  #mafless0.01   0
Het[less_line1,]

Exclu <- Het_Aut_imiss[c(more_line2, less_line1),]
dim(Exclu)
Exclu

plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, main = "Autosomal heter
ozygosity and call rate", sub = paste(Cohort, maf, sep=" ") ,xlab="% Heterozy
gosity", ylab="F_MISS", , col = rgb(0.2, 0.8, 0.5, alpha = 0.5), pch = 19)
```

```
temp <- subset(Exclu)

plot(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, main = "Autosomal heter
ozygosity and call rate", sub = paste(Cohort, maf, sep=" ") ,xlab="% Heterozy
gosity", ylab="F_MISS", col = rgb(0.2, 0.8, 0.5, alpha = 0.5), pch = 19)

temp <- subset(Exclu)

points(temp$percent_het, temp$F_MISS, col="red", pch=16, cex=0.8)
legend("topright", c(">3xSD"), fill=c("red"))

identify(Het_Aut_imiss$percent_het, Het_Aut_imiss$F_MISS, labels=Het_Aut_imis
s$ID)
```

Identify duplicated and related samples

Relatedness identifies possible cryptic relationships between individuals thought to be unrelated, sample identity errors in family data, improperly duplicated samples and contaminated samples that have escaped other screenings (O. Igo et al.,2016).

Mendel's law (1866) for diploid organisms say that at any given locus, each individual has two genes, one maternal and one paternal. Thus, individuals carry at a locus pieces of DNA that are copied through repeated segregations from their ancestors. Relatives who share a common ancestor may both carry copies of the sample ancestral piece of DNA. Such cases said to be Identical by Descent (IBD).

IBD has three coefficients (Z0, Z1, Z2) which are automatically produced when running –genome command on PLINK and estimate the probability two individuals sharing 0, 1 or 2 alleles. Using this probability model and method we estimate the Pr(Z) for any given marker. Some pedigree errors can be corrected by consulting original records, while other are corrected based on the inferred genetic relationship (Laurie et al., 2010).

Identify by state, IBS, is used to calculate for each pair of individuals the average proportion of alleles shared in common at genotyped SNPs (including only the autosomal regions). Because this method works better when only independent SNPs are included, complex regions of extended linkage disequilibrium (LD) are removed from the merged dataset (Purcell S, et al. 2007). The remaining regions are pruned so that no pair of SNPs within a given region is correlated (r2>0.2).

We want keep individuals that have less than 2nd degree of relatedness (pi_hat < 0.2), but first we subsetting out the duplicates with pi_hat > 0.9 where we find three individuals, apply pair-wise (0.2 < pi_hat < 0.9)

We we apply pi_hat treshold, a decision needs to me made based on which sample we keep and which we remove.

The most common ways of choosing are:

- Based on the lowest call rate between the duplicated/related 2 samples, I want to keep the one with the highest
- When I have cases and controls I keep the sample from the cases
- Another hint is when I have 3 samples, lets say 1,2,3 and the 1 is related with 2 & 2 is related with 3, it makes more sense to remove sample 2, so I can save 1 sample.

```
dir <- "…/"
Cohort <- " Firstmerge_all_Cohorts "
file <- "_prefiltered-chr1-22-"
maf <- "mafgte0.01"

genome <- read.table(paste(dir, "tab-", Cohort, file, maf,  "-noCR-LDpruned0.
2.genome.genome", sep =""), sep = "\t", header=TRUE, fill=TRUE)

summary(genome$PI_HAT)

which(is.na(genome$PI_HAT))

sorted_PI_HAT <- sort(genome$PI_HAT) #Plotting the PIHATs in rank order -
plot(sorted_PI_HAT, main= Cohort, xlab = "rank order of PI_HAT", ylab = "PI_H
AT")
abline(h = 0.2, lwd = 2, col = 2, lty = 2)

# subsetting out the duplicates (PH_HAT > 0.9)

wPiHatmore0.9 <- which(genome$PI_HAT > 0.9)
length(wPiHatmore0.9) #3
genome[c (2840, 5216, 5418),]

wPiHatmore0.2 <- which(genome$PI_HAT > 0.2 & genome$PI_HAT < 0.9 )
length(wPiHatmore0.2) #2
genome [ c(17240, 21008),]
```

```
write.table(Aut_imiss, "find_duplicates_call_rates.txt", sep="\t", col.names=
TRUE, row.names=FALSE, quote=FALSE)
```

- Choose which pair of duplicated/related individuals we keep:

SAMPLE 1	SAMPLE 2	Missingness Sample 1	Missingness Sample 2	Decision
urn:wtsi:433501_A04_OAfg6308476	urn:wtsi:493153_G01_OAfg6308476	0.0005096	0.002091	keep sample 2
urn:wtsi:433501_G02_OAfg6308466	urn:wtsi:493153_E01_OAfg6308466	0.0003628	0.001834	keep sample 2
urn:wtsi:433501_G03_OAfg6308474	urn:wtsi:493153_F01_OAfg6308474	0.0003302	0.001296	keep sample 2
urn:wtsi:493153_H10_fungenQQ6828557	urn:wtsi:533231_E04_fungenQQ7246857	0.002658	0.0002731	keep sample 1
urn:wtsi:533231_D01_fungenQQ7246832	urn:wtsi:533232_H01_fungenQQ7442739	0.001054	0.001944	keep sample 1

Ethnicity outliers (PCA) - Merged with 1000 Genome Reference panel

```
dir <- "…/ETHNICITY/"
Cohort <- "Firstmerge_all_Cohorts_merged_1kg"
chip <- "CoreExome"

genome <-read.table("/…/ETHNICITY/Firstmerge_all_Cohorts_merged_1kg-mafgte0.0
1-noCR-LDpruned0.2.genome.mds.mds",  header=TRUE, stringsAsFactors=FALSE)
pop_info <- read.table("/…/ETHNICITY/PopulationFile_ Firstmerge_all_Cohorts_1
kg.txt", sep="\t", header=TRUE)


genome_popinfo <- merge(genome, pop_info, by.x = "FID", by.y = "ID", all.x =
TRUE)
dim(genome_popinfo)
names(genome_popinfo)
head(genome_popinfo)
Graph_title <- paste(Cohort,chip, "vs 1000genomes", sep=" ")
summary(genome_popinfo$Population)

# 1rst plot:
plot(genome_popinfo$C1, genome_popinfo$C2, col="grey", main= Graph_title, xla
b="Component 1", ylab="Component 2") #just to check everything is plotted

temp <- subset(genome_popinfo, genome_popinfo$Population=="FunGen1+2+QQ+QQ2_N
ohips")        #Fungen1+2+QQ+QQ2
points(temp$C1, temp$C2, col="red", pch=16, cex=0.8)

legend("topright", c("FunGen1+2+QQ+QQ2_Nohips"), pch=c(16), col=c("red"),cex=
0.6)

identify(genome_popinfo$C1, genome_popinfo$C2, labels=genome_popinfo$FID, cex
=0.6)

# 2nd plot:


plot(genome_popinfo$C1, genome_popinfo$C2, col="grey", main= Graph_title, xla
b="Component 1", ylab="Component 2") #just to check everything is plotted
```

```
# HapMap African ancestry individuals from SW US  n=61
temp <- subset(genome_popinfo, genome_popinfo$Population=="ASW")

# (LWK) Luhya individuals n=97
points(temp$C1, temp$C2, col="darkolivegreen3", pch=2)
temp <- subset(genome_popinfo, genome_popinfo$Population=="LWK")

# (YRI) Yoruba individuals n=88
points(temp$C1, temp$C2, col="forestgreen")
temp <- subset(genome_popinfo, genome_popinfo$Population=="YRI")

# (CHB) Han Chinese in Beijing n=97
points(temp$C1, temp$C2, col="chartreuse3")
temp <- subset(genome_popinfo, genome_popinfo$Population=="CHB")

# (CHS) Han Chinese South  n=100
points(temp$C1, temp$C2, col="darkblue")
temp <- subset(genome_popinfo, genome_popinfo$Population=="CHS")

# JPT Japanese individuals n=89
points(temp$C1, temp$C2, col="deepskyblue")
temp <- subset(genome_popinfo, genome_popinfo$Population=="JPT")

# Colombian in Medellin, Colombia  n=60
points(temp$C1, temp$C2, col="darkslategray1")
temp <- subset(genome_popinfo, genome_popinfo$Population=="CLM")

# Puerto Rican in Puerto Rico n=55
points(temp$C1, temp$C2, col="mediumpurple4")
temp <- subset(genome_popinfo, genome_popinfo$Population=="PUR")

# HapMap Mexican individuals from LA California n=66
points(temp$C1, temp$C2, col="mediumpurple1")
temp <- subset(genome_popinfo, genome_popinfo$Population=="MXL")

# CEPH individuals  n=85
points(temp$C1, temp$C2, col="mediumorchid1")
temp <- subset(genome_popinfo, genome_popinfo$Population=="CEU")

# Toscan individuals n=98
points(temp$C1, temp$C2, col="mediumvioletred")
temp <- subset(genome_popinfo, genome_popinfo$Population=="TSI")

points(temp$C1, temp$C2, col="yellow")
```

```
# HapMap Finnish individuals from Finland n=93
temp <- subset(genome_popinfo, genome_popinfo$Population=="FIN")

points(temp$C1, temp$C2, col="pink1")

# Iberian populations in Spain n=14 n=93
temp <- subset(genome_popinfo, genome_popinfo$Population=="IBS")

points(temp$C1, temp$C2, col="orange")
# British individuals from England and Scotland (GBR) n=89

temp <- subset(genome_popinfo, genome_popinfo$Population=="GBR")

points(temp$C1, temp$C2, col="lightsalmon3")

# Firstmerge_all_Cohorts
temp <- subset(genome_popinfo, genome_popinfo$Population=="Firstmerge_all_Coh
orts")

points(temp$C1, temp$C2, col="red", pch=16, cex=0.8)

legend("bottomright", c("ASW","LWK","YRI","CHB","CHS","JPT","CLM","PUR","MXL"
,"CEU","TSI","FIN","IBS","GBR","FunGen1+2+QQ+QQ2_Nohips"), pch=c(2,1,1,1,1,1,
1,1,1,1,1,1,1,1,16), col=c("darkolivegreen3","forestgreen","chartreuse3","dar
kblue","deepskyblue","darkslategray1","mediumpurple4","mediumpurple1","medium
orchid1","mediumvioletred","yellow","pink1","orange","lightsalmon3","red"),ce
x=0.6)

identify(genome_popinfo$C1, genome_popinfo$C2, labels=genome_popinfo$FID, cex
=0.6)


# Zoom in

#Plotting with axis changed to zoom into the European section:
plot(genome_popinfo$C1, genome_popinfo$C2, col="grey", main= Graph_title, xla
b="Component 1", ylab="Component 2") #just to check everything is plotted


plot(genome_popinfo$C1, genome_popinfo$C2, col="white", main= Graph_title , x
lab="Component 1", ylab="Component 2", xlim =c(-0.025,-0.005), ylim =c(0.015,
0.045)) #change to white for plot
```

```
# HapMap African ancestry individuals from SW US n=61
temp <- subset(genome_popinfo, genome_popinfo$Population=="ASW")
points(temp$C1, temp$C2, col="darkolivegreen3", pch=2)


# (LWK) Luhya individuals n=97

temp <- subset(genome_popinfo, genome_popinfo$Population=="LWK")


points(temp$C1, temp$C2, col="forestgreen")


# (YRI) Yoruba individuals n=88

temp <- subset(genome_popinfo, genome_popinfo$Population=="YRI")

points(temp$C1, temp$C2, col="chartreuse3")


# (CHB) Han Chinese in Beijing n=97
temp <- subset(genome_popinfo, genome_popinfo$Population=="CHB")
points(temp$C1, temp$C2, col="darkblue")


# (CHS) Han Chinese South  n=100
temp <- subset(genome_popinfo, genome_popinfo$Population=="CHS")
points(temp$C1, temp$C2, col="deepskyblue")


# JPT Japanese individuals n=89
temp <- subset(genome_popinfo, genome_popinfo$Population=="JPT")
points(temp$C1, temp$C2, col="darkslategray1")

# Colombian in Medellin, Colombia  n=60
temp <- subset(genome_popinfo, genome_popinfo$Population=="CLM")
points(temp$C1, temp$C2, col="mediumpurple4")

# Puerto Rican in Puerto Rico n=55
temp <- subset(genome_popinfo, genome_popinfo$Population=="PUR")
points(temp$C1, temp$C2, col="mediumpurple1")
```

```
# HapMap Mexican individuals from LA California n=66
temp <- subset(genome_popinfo, genome_popinfo$Population=="MXL")
points(temp$C1, temp$C2, col="mediumorchid1")

# CEPH individuals  n=85
temp <- subset(genome_popinfo, genome_popinfo$Population=="CEU")
points(temp$C1, temp$C2, col="mediumvioletred")

# Toscan individuals n=98
temp <- subset(genome_popinfo, genome_popinfo$Population=="TSI")
points(temp$C1, temp$C2, col="yellow")

# HapMap Finnish individuals from Finland n=93
temp <- subset(genome_popinfo, genome_popinfo$Population=="FIN")
points(temp$C1, temp$C2, col="pink1")

# Iberian populations in Spain n=14 n=93
temp <- subset(genome_popinfo, genome_popinfo$Population=="IBS")
points(temp$C1, temp$C2, col="orange")

# British individuals from England and Scotland (GBR) n=89
temp <- subset(genome_popinfo, genome_popinfo$Population=="GBR")
points(temp$C1, temp$C2, col="lightsalmon3")

#FunGen samples n=220
temp <- subset(genome_popinfo, genome_popinfo$Population=="FunGen1+2+QQ+QQ2_N
ohips")
points(temp$C1, temp$C2, col="red", pch=16, cex=0.8)



legend("topright", c("ASW","LWK","YRI","CHB","CHS","JPT","CLM","PUR","MXL","C
EU","TSI","FIN","IBS","GBR","FunGen1+2+QQ+QQ2_Nohips"), pch=c(2,1,1,1,1,1,1,1
,1,1,1,1,1,1,16), col=c("darkolivegreen3","forestgreen","chartreuse3","darkbl
ue","deepskyblue","darkslategray1","mediumpurple4","mediumpurple1","mediumorc
hid1","mediumvioletred","yellow","pink1","orange","lightsalmon3","red"),cex=0
.6)

identify(genome_popinfo$C1, genome_popinfo$C2, labels=genome_popinfo$FID, cex
=0.6)
```

Exclusion Summary

- 9 Individuals to exclude after SAMPLE QC, saved as "9.Exclusions.Sample.QC.txt" :

urn:wtsi:433501_D03_OAfg6308471	urn:wtsi:433501_D03_OAfg6308471
urn:wtsi:433501_B03_OAfg6308469	urn:wtsi:433501_B03_OAfg6308469
urn:wtsi:533231_G09_fungenQQ7246899	urn:wtsi:533231_G09_fungenQQ7246899
urn:wtsi:533231_C08_fungenQQ7246887	urn:wtsi:533231_C08_fungenQQ7246887
urn:wtsi:493153_G01_OAfg6308476	urn:wtsi:493153_G01_OAfg6308476
urn:wtsi:493153_E01_OAfg6308466	urn:wtsi:493153_E01_OAfg6308466
urn:wtsi:493153_F01_OAfg6308474	urn:wtsi:493153_F01_OAfg6308474
urn:wtsi:493153_H10_fungenQQ6828557	urn:wtsi:493153_H10_fungenQQ6828557
urn:wtsi:533231_D01_fungenQQ7246832	urn:wtsi:533231_D01_fungenQQ7246832

- Remove them to continue with SNP QC

```
plink --bfile Firstmerge_all_Cohorts --remove 9.Exclusions.Sample.QC.txt --ma
ke-bed --out Firstmerge_all_Cohorts
```

SNP QC Steps taken

We perform the SNP QC seperately for

- Autosome region: Chr1-22
- Chromosome 23: X-chr nonPar is performed in females only
- Chromosome 25: X-chr PAR ½ QC is performed in males and females

And then we apply for each region:

1. SNP Call rate 98%
2. HWE p_value < 1e-4

We use the hh file after Sample and SNP QC, to Exclude the nonPAR x-Chr SNPS that have haploid male genotypes.

Run SNP-QC Script

I have saved the following commands in perl script named "Produce_snpqc_summary_stats_Gencall.sh", producing all summary files SNP QC requires:

```
#---------- Produce_snpqc_summary_stats_Gencall.sh --------------
#!/bin/sh

# A script to run through the PLINK steps for Core Exome Chip using the GenCa
lled data that has had the samples excluded from Sample-QC

# Required parameters:
# Directory
# File name

#
DIR=$1
FILE=$2

echo File used here is $DIR/$FILE
#======================#
# 1. Autosomal (chr1-22)
#======================#
plink --noweb --bfile $DIR/$FILE  --autosome --make-bed --out $DIR/$FILE-chr1
```

```
-22

## Call rate ##
plink --noweb --bfile $DIR/$FILE-chr1-22  --missing --out $DIR/$FILE-chr1-22-
missing

# Produce a log file giving samples excluded at CR 98% to check against R res
ult = Gencall threshold
plink --noweb --bfile $DIR/$FILE-chr1-22 --geno 0.02  --out $DIR/$FILE-chr1-2
2-geno0.02

## HWE ##
plink --noweb --bfile $DIR/$FILE-chr1-22 --hardy  --out $DIR/$FILE-chr1-22-ha
rdy

#Produce a log file giving the number of variants excluded with HWE P<1E-4
plink --noweb --bfile $DIR/$FILE-chr1-22 --hwe 0.0001  --out $DIR/$FILE-chr1-
22-hwe0.0001


# MAF ##
plink --noweb --bfile $DIR/$FILE-chr1-22 --freq  --out $DIR/$FILE-chr1-22-fre
q

#Filter the file and produce a SNP list that can be used to check the resulti
ng files:

plink --noweb --bfile $DIR/$FILE-chr1-22  --geno 0.02 --make-bed --out $DIR/$
FILE-chr1-22-geno0.02
plink --noweb --bfile $DIR/$FILE-chr1-22-geno0.02  --hwe 0.0001 --make-bed --
out $DIR/$FILE-chr1-22-geno0.02-hwep1e-4
plink --noweb --bfile $DIR/$FILE-chr1-22-geno0.02-hwep1e-4 --write-snplist --
out $DIR/$FILE-chr1-22-geno0.02-hwep1e-4-GENCALL-SNP-INCLUSIONS
rm $DIR/$FILE-chr1-22-geno0.02.bed
rm $DIR/$FILE-chr1-22-geno0.02.bim
rm $DIR/$FILE-chr1-22-geno0.02.fam
rm $DIR/$FILE-chr1-22-geno0.02-hwep1e-4.bed
rm $DIR/$FILE-chr1-22-geno0.02-hwep1e-4.bim
rm $DIR/$FILE-chr1-22-geno0.02-hwep1e-4.fam

#==================#
#  2. X-chr PAR1/2
#==================#
```

```
plink --noweb --bfile $DIR/$FILE  --chr 25 --make-bed --out $DIR/$FILE-chrXPA
R1-2

## Call rate ##
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2 --missing --out $DIR/$FILE-chrXPA
R1-2-missing

# Produce a log file giving samples excluded at CR 98% to check against R res
ult = Gencall threshold
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2 --geno 0.02  --out $DIR/$FILE-chr
XPAR1-2-geno0.02

## HWE ##
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2 --hardy  --out $DIR/$FILE-chrXPAR
1-2-hardy

#Produce a log file giving the number of variants excluded with HWE P<1E-4
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2 --hwe 0.0001  --out $DIR/$FILE-ch
rXPAR1-2-hwe0.0001


# MAF ##
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2 --freq  --out $DIR/$FILE-chrXPAR1
-2-freq

#Filter the file and produce a SNP list that can be used to check the resulti
ng files:

plink --noweb --bfile $DIR/$FILE-chrXPAR1-2  --geno 0.02 --make-bed --out $DI
R/$FILE-chrXPAR1-2-geno0.02
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2-geno0.02  --hwe 0.0001 --make-bed
--out $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4
plink --noweb --bfile $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4 --write-snplist
--out $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4-GENCALL-SNP-INCLUSIONS
rm $DIR/$FILE-chrXPAR1-2-geno0.02.bed
rm $DIR/$FILE-chrXPAR1-2-geno0.02.bim
rm $DIR/$FILE-chrXPAR1-2-geno0.02.fam
rm $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4.bed
rm $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4.bim
rm $DIR/$FILE-chrXPAR1-2-geno0.02-hwep1e-4.fam
```

```
#==================================#
#  3. X-chr nonPAR in femlaes only
#==================================#

plink --noweb --bfile $DIR/$FILE  --chr 23 --filter-females --make-bed --out
$DIR/$FILE-chrX-femalesonly

## Call rate ##
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly  --missing --out $DIR/$FILE
-chrX-femalesonly-missing

# Produce a log file giving samples excluded at CR 98% to check against R res
ult = Gencall threshold
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly --geno 0.02  --out $DIR/$FI
LE-chrX-femalesonly-geno0.02

## HWE ##
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly --hardy  --out $DIR/$FILE-c
hrX-femalesonly-hardy

#Produce a log file giving the number of variants excluded with HWE P<1E-4
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly --hwe 0.0001  --out $DIR/$F
ILE-chrX-femalesonly-hwe0.0001


# MAF ##
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly --freq  --out $DIR/$FILE-ch
rX-femalesonly-freq

#Filter the file and produce a SNP list that can be used to check the resulti
ng files:

plink --noweb --bfile $DIR/$FILE-chrX-femalesonly  --geno 0.02 --make-bed --o
ut $DIR/$FILE-chrX-femalesonly-geno0.02
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly-geno0.02  --hwe 0.0001 --ma
ke-bed --out $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4
plink --noweb --bfile $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4 --write-s
nplist --out $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4-GENCALL-SNP-INCLUS
IONS
rm $DIR/$FILE-chrX-femalesonly-geno0.02.bed
rm $DIR/$FILE-chrX-femalesonly-geno0.02.bim
rm $DIR/$FILE-chrX-femalesonly-geno0.02.fam
rm $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4.bed
```

```
rm $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4.bim
rm $DIR/$FILE-chrX-femalesonly-geno0.02-hwep1e-4.fam


#------------ END ------------------#
```

After this step, I switch to Jupyter Notebook and I load my files in R as follows:

Autosomal_SNPS_1-22

```
file <- " Firstmerge_all_Cohorts.samplesout-chr1-22"
dir <- "/…/4_1_2_SNP_QC/"

hardy <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chr1-22-hardy.hwe",
header=TRUE, fill=TRUE)
freq <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chr1-22-freq.frq", h
eader=TRUE, fill=TRUE)
lmiss <- read.table("/Firstmerge_all_Cohorts.samplesout-chr1-22-missing.lmiss
", header=TRUE, fill=TRUE)

head(hardy)
head(freq)
head(lmiss)

names(hardy)
names(freq)
names(lmiss)

dim(hardy)
dim(freq)
dim(lmiss)

hardy <- hardy[st1,]
dim(hardy)

hardy_freq <- merge(hardy,freq, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FALS
E)
head(hardy_freq)
tail(hardy_freq)
```

```
names(hardy_freq)
dim(hardy_freq)

#Get rid of duplicated columns
hardy_freq$CHR.y <- NULL
hardy_freq$A1.y <- NULL
hardy_freq$A2.y <- NULL

names(hardy_freq)
data <- merge(hardy_freq, lmiss, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FAL
SE)
head(data)
tail(data)

names(data)
dim(data)

data$CHR <- NULL
data$TEST<- NULL

names(data) <- c("SNP","CHR","A1","A2","GENO","O.HET.","E.HET.", "HWE_P", "MA
F","NCHROBS","N_MISS","N_GENO","F_MISS")
head(data)
tail(data)
names(data)
dim(data)

Finaldata <- data

head(Finaldata)
tail(Finaldata)
names(Finaldata)
dim(Finaldata)

hist(lmiss$F_MISS, freq=TRUE, col="light blue", border= "black", main = "SNP
Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS", ylab = "Frequenc
y", ylim=c(0,550), xlim=c(0,1.2))

write.table(Finaldata, paste(dir,file, "-GENCALL-SUMMARY_STATS", sep=""), sep
="\t", col.names=TRUE, row.names=FALSE, quote=FALSE)
```

```
exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)
length(exclusions1)


crexclus <- which(Finaldata$F_MISS > 0.02)
length(crexclus) #9906

hweexclus <- which(Finaldata$HWE_P <0.0001)
length(hweexclus) #85

hwecrexclus <- which(data$HWE_P <0.0001 | data$F_MISS > 0.02)
length(hwecrexclus) #9948

exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)

write.table(exclusions1, paste(dir,file, "-GENCALL-CALLRATE0.98-HWEP1e-4-EXCL
USIONS", sep=""), sep="\t", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

XPAR_NonPAR_SNPs

```
file <- " Firstmerge_all_Cohorts.samplesout-chrX-femalesonly"
dir <- "/…/"


hardy <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chrXPAR1-2-hardy.hw
e", header=TRUE, fill=TRUE)

freq <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chrXPAR1-2-freq.frq"
, header=TRUE, fill=TRUE)

lmiss <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chrXPAR1-2-missing.
lmiss", header=TRUE, fill=TRUE)

head(hardy)
head(freq)
```

```
head(lmiss)

names(hardy)
names(freq)
names(lmiss)

dim(hardy)
dim(freq)
dim(lmiss)

st1 <- which(hardy$TEST =="ALL(NP)")
length(st1) #155

hardy <- hardy[st1,]
dim(hardy) # 155 9

hardy_freq <- merge(hardy,freq, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FALS
E)
head(hardy_freq)
tail(hardy_freq)

names(hardy_freq)
dim(hardy_freq)

#Get rid of duplicated columns
hardy_freq$CHR.y <- NULL
hardy_freq$A1.y <- NULL
hardy_freq$A2.y <- NULL

names(hardy_freq)
data <- merge(hardy_freq, lmiss, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FAL
SE)
head(data)
tail(data)

data$CHR <- NULL
data$TEST<- NULL

names(data) <- c("SNP","CHR","A1","A2","GENO","O.HET.","E.HET.", "HWE_P", "MA
F","NCHROBS","N_MISS","N_GENO","F_MISS")
head(data)
tail(data)
```

```
names(data)
dim(data)

Finaldata <- data

head(Finaldata)
tail(Finaldata)
names(Finaldata)
dim(Finaldata)

hist(lmiss$F_MISS, freq=TRUE, col="light blue", border= "black", main = "SNP
Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS", ylab = "Frequenc
y", ylim=c(0,550), xlim=c(0,1.2))

write.table(Finaldata, paste(dir,file, "-GENCALL-SUMMARY_STATS", sep=""), sep
="\t", col.names=TRUE, row.names=FALSE, quote=FALSE)



exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)
length(exclusions1)


crexclus <- which(Finaldata$F_MISS > 0.02)
length(crexclus) #9906

hweexclus <- which(Finaldata$HWE_P <0.0001)
length(hweexclus) #85

hwecrexclus <- which(data$HWE_P <0.0001 | data$F_MISS > 0.02)
length(hwecrexclus) #9948

exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)

write.table(exclusions1, paste(dir,file, "-GENCALL-CALLRATE0.98-HWEP1e-4-EXCL
USIONS", sep=""), sep="\t", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

XPAR1/2 SNPs

- XPAR SNPs

```
file <- " Firstmerge_all_Cohorts.samplesout-chrX-femalesonly"

dir <- "/…/"


hardy <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chrX-femalesonly-ha
rdy.hwe", header=TRUE, fill=TRUE)
freq <- read.table("/…/Firstmerge_all_Cohorts.samplesout-chrX-femalesonly-fre
q.frq", header=TRUE, fill=TRUE)
lmiss <- read.table("/…/ Firstmerge_all_Cohorts.9.samplesout-chrX-femalesonly
-missing.lmiss", header=TRUE, fill=TRUE)

head(hardy)
head(freq)
head(lmiss)

names(hardy)
names(freq)
names(lmiss)

dim(hardy)
dim(freq)
dim(lmiss)

st1 <- which(hardy$TEST =="ALL(NP)")
length(st1) #155

hardy <- hardy[st1,]
dim(hardy) # 155 9

hardy_freq <- merge(hardy,freq, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FALS
E)
```

```
head(hardy_freq)
tail(hardy_freq)

names(hardy_freq)
dim(hardy_freq)

#Get rid of duplicated columns
hardy_freq$CHR.y <- NULL
hardy_freq$A1.y <- NULL
hardy_freq$A2.y <- NULL

names(hardy_freq)
data <- merge(hardy_freq, lmiss, by.x="SNP", by.y="SNP", all.x=TRUE, sort=FAL
SE)
head(data)
tail(data)

data$CHR <- NULL
data$TEST<- NULL

names(data) <- c("SNP","CHR","A1","A2","GENO","O.HET.","E.HET.", "HWE_P", "MA
F","NCHROBS","N_MISS","N_GENO","F_MISS")
head(data)
tail(data)
names(data)
dim(data)

Finaldata <- data

head(Finaldata)
tail(Finaldata)
names(Finaldata)
dim(Finaldata)

hist(lmiss$F_MISS, freq=TRUE, col="light blue", border= "black", main = "SNP
Call Rate", sub = "Fungen1-Fungen2-QQ-QQ2", xlab = "F_MISS", ylab = "Frequenc
y", ylim=c(0,550), xlim=c(0,1.2))

write.table(Finaldata, paste(dir,file, "-GENCALL-SUMMARY_STATS", sep=""), sep
="\t", col.names=TRUE, row.names=FALSE, quote=FALSE)
```

```
exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)
length(exclusions1)


crexclus <- which(Finaldata$F_MISS > 0.02)
length(crexclus) #9906

hweexclus <- which(Finaldata$HWE_P <0.0001)
length(hweexclus) #85

hwecrexclus <- which(data$HWE_P <0.0001 | data$F_MISS > 0.02)
length(hwecrexclus) #9948

exclusions1 <- Finaldata[hwecrexclus, 1]
head(exclusions1)

write.table(exclusions1, paste(dir,file, "-GENCALL-CALLRATE0.98-HWEP1e-4-EXCL
USIONS", sep=""), sep="\t", col.names=FALSE, row.names=FALSE, quote=FALSE)
```

Produce a file for the X-chr nonPAR

```
hh <- read.table("/…/Firstmerge_all_Cohorts.Individual.SNP.QCed.hh", header=F
ALSE, fill=TRUE, stringsAsFactors=FALSE)
head(hh)
tail(hh)
dim(hh)
names(hh)
hhtable <- as.data.frame(table(hh$V3))
head(hhtable)

names(hhtable) <- c("SNP","Freq")
dim(hhtable)
plot(hhtable$Freq)
hist(hhtable$Freq)

write.table(hhtable, "hh-table.xchr.after.SNP.Sample.QC.txt1", sep="\t", col.
```

```
names=TRUE, row.names=FALSE, quote=FALSE)
write.table(hhtable[,1], "hh-table.xchr.after.SNP.Sample.QC.txt2", sep="\t",
col.names=FALSE, row.names=FALSE, quote=FALSE)
```