UGLI2 (release 1.0) Quality Control Report

The University Medical Center of Groningen Genetics Lifelines Initiative (UGLI) is a project that intends to genotype all volunteers of the Lifelines project. This report summarizes the quality control (QC) process of the first release of UGLI comprising the genotype of 29,166 participants assessed using the FinnGen Thermo Fisher Axiom® custom array. In this QC screening we included all genotyped samples, but we focused on QC of genetic markers on the autosomes and chromosomes X (N=617,715 and 22,405 markers, respectively).

In brief, first sample specific priors for the genotype calling algorithm were generated using the first 25 plates, that seemed to have performed well, using the tool *simple_ssp* tool provided by Thermo Fisher. Next the genotypes were called using the Axiom Analysis Suite developed by Thermo Fisher. Then the genotypes were exported using the long format tool (*0.apt2-format-long_UMCG.sh*) from Thermo Fisher and finally they were converted to binary PLINK format to perform the QC. This started by first checking concordance of duplicate markers and samples. Then the data were filtered for low quality samples and markers with a two-steps procedure of call rate thresholding. Further possible genotyping errors were assessed (i) at the marker level by detecting variants with a very low minor allele frequency and that deviated very significantly from Hardy-Weinberg equilibrium (HWE); and (ii) at the sample level by evaluating heterozygosity. We then evaluated samples mix-ups in two levels: i) concordance of reported sex with sex derived from genotyping data from the X chromosome, and ii) concordance of reported family information (Lifelines pedigree) and thus of the expected genome sharing between relatives with the observed sharing from genotyped data (genetic kinship). For this latter check also genotype data from Lifelines samples genotyped used two previous genotyping chips (CytoSNP 250k and the Infinium Global Screening Array® (GSA) MultiEthnic Disease Version) were used. Subsequently, we ascertained Mendelian errors and further removed genetic markers that deviated from HW in unrelated individuals. Finally, population stratification was inspected by a principle components analysis (PCA), incorporating samples from the 1000 Genomes (1000G) project. These summarized steps are shown in **Figure 1**, where each step is annotated together with the required input and whether the step generates a graphical output or a report.

Step-wise quality control

1. Pre-quality control steps

Genotype calls of the autosomal, pseudo-autosomal chromosome XY, chromosome Y and mitochondrial (MT) genetic variants were determined from Affymetrix CEL files using Thermofisher's Axiom Analysis Suite. The genotypes were called in batches of 12 plates (n=952 samples). Opticall (https://opticall.bitbucket.io) was used to call the genotypes of the genetic variants in the non-pseudoautosomal regions of the X chromosome.

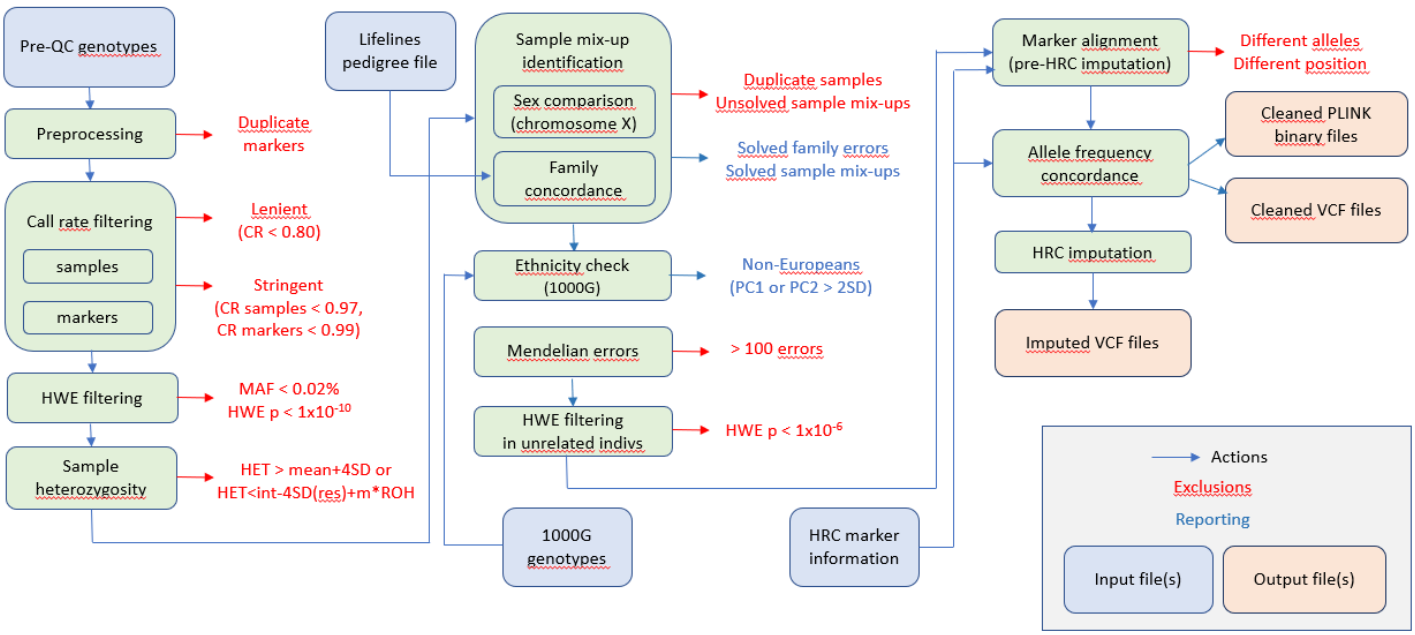

Figure 1. Steps and metrics evaluated in the quality control of the UGLI2 genotype data.

Assigned genotypes were next converted to PLINK (https://www.cog-genomics.org/plink/1.9/) binary files and separated into chromosomes (autosomal 1-22, X, Y, XY, and MT) to be further processed. For the remainder of the quality control only the autosomal markers and the markers from chromosome X (including the pseudoautosomal regions) were checked, thus the 598 markers from chromosome Y and the 521 mitochondrial markers were excluded.

2. Filtering duplicate markers and samples

For autosomal and pseudo-autosomal chromosome we removed duplicate markers and samples. For this, marker names were converted to chr_pos_A1_A2 ids, where A1 and A2 are the two alleles in alphabetic order. In this way tri-allelic markers make two different markers as well as single nucleotide polymorphisms (SNPs) and insertion-deletion polymorphisms (indels) at the same position. Duplicate markers were identified and got an additional identifier ":1", ":2", etc. attached to their name. Next separate subsets of markers were created based on these identifiers with the PLINK v1.9b3.32 command *–extract*. Before merging these subsets of markers, the duplicate marker identifiers were removed again and genotype concordance was checked with the command *–merge-mode 7*. If more than 1% of the calls was discordant, both markers were excluded with the *–exclude* command. For the remainder of the duplicate markers, which proved to be concordant, the call rate was calculate with the *–missing* command. Next, the marker with lowest call rate was identified and removed. As a final step all additional identifiers for the duplicate markers (i.e. ":1", ":2", etc.) were removed from the marker names. For built-in duplicate samples a similar approach was followed. The genomic relation between samples was not checked at this time, implying that unintended duplicate samples (or monozygotic twins) were not considered in this step.

We identified and removed 2736 duplicated (by position and allele) markers and 198 duplicate samples.

3. Filtering markers and samples with a low call rate

Autosomal and pseudo-autosomal markers with high missing rate were removed using a two-thresholds two-steps process: first by samples and then by markers, filtering first with a lenient missing rate threshold (20%) and then by applying a more stringent missing rate threshold (1% for markers and 3% for samples, per suggestion ThermoFisher). All the steps here were done *--missing* --*remove* and --*exclude*, following this workflow: 2a. Calculate missing rate per sample and remove samples with missing rate >20%; 2b. Calculate missing rate for markers and remove markers with missing rate >20%; 2c. Recalculate missing rate for samples and remove samples with missing rate >3%; 2d. Recalculate the missing rate for markers and remove markers with missing rate >1%.

After the lenient call rate filter (80%, i.e. missing rate=20%) (excluding 172 markers and 1 sample), the distributions of call rates are very skewed as expected (**Figure 2**). As advised by Thermo Fisher, we decided for a stringent sample call rate threshold of 97% and a marker call rate of 99%. After call rate filtering 27,801 (99.4%) samples and 593,575 (96.5%) markers remained.

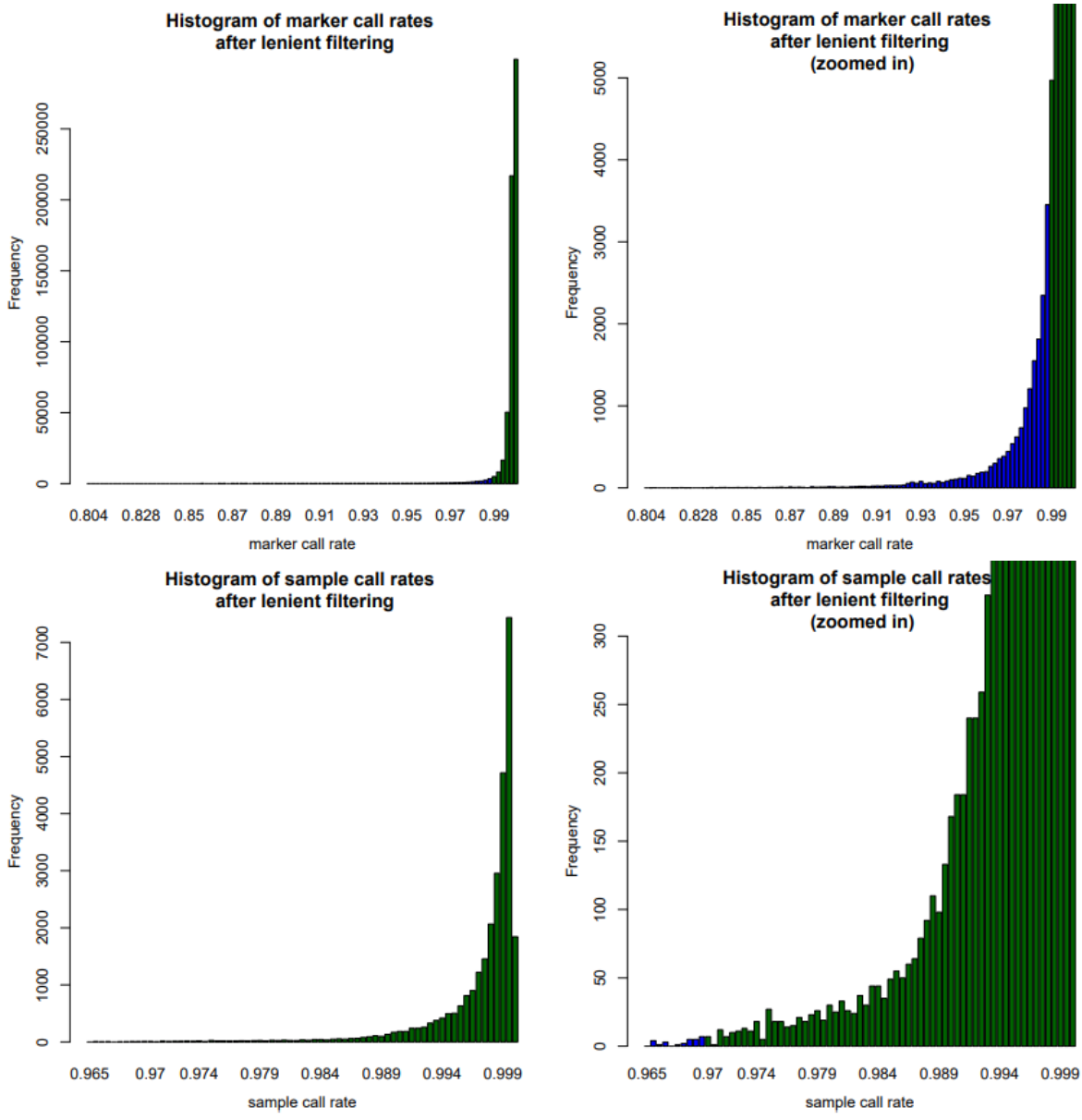

Figure 2: Distribution of the call rates after lenient filtering. The bottom graphs represent the marker call rates; the bottom one the sample call rates. The bars are colored blue in case of a marker call rate <=99% or a sample call rate <=97% and green for a marker call rate >99% or a sample call rate >97%.

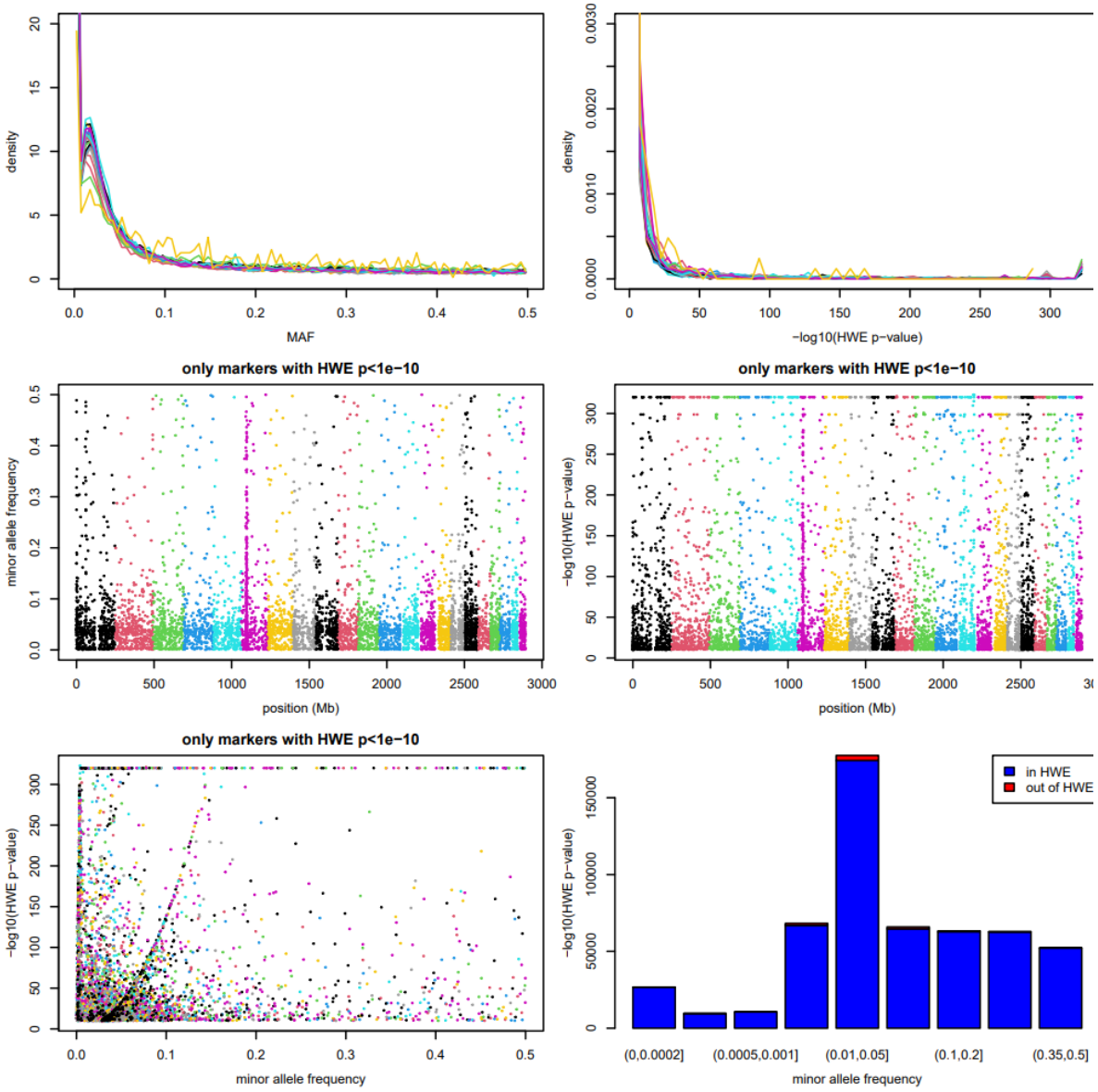

Figure 3: Various graphs showing the distribution of and relation between minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) p-value per chromosome. Upper left and upper right plot show the distributions of the MAF and HWE p-value, respectively. The middle plots show the MAF (left) and -log10(HWE p-value) (right) of the markers with a HWE p-value <1x10^{-10} distributed over the genome. Lower left plot shows the relation between MAF and -log10(HWE p-value). Lower right plots shows the number of markers with a HWE p-value >1x10^{-10} (blue) and <1x10^{-10} (red) per MAF bin.

4. Minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE)

We calculated the allele frequencies and HWE p-values using PLINK commands `--freq` and `--hardy`. Markers with a minor allele frequency (MAF) <0.02% and/or markers with a HWE p-value <1x10^{-10} were considered uninformative and of poor quality. No clear relation was observed

between MAF and HWE p-value (**Figure 3**). For the HWE test no pedigree information was available yet, so a lenient threshold is used. This HWE QC step is repeated after establishing family relations of all samples (see step 9).

A total of 83,195 (14.0%) markers were found to have a MAF below the threshold (of which 56,467 are monomorphic) and another 7,163 (1.4%) were out of HWE. These were removed in this step.

5. Sample heterozygosity

A common step in quality control of genome-wide arrays is to check for sample heterozygosity. Outliers showing excess or depletion in heterozygotes genotypes may be due to DNA contamination or issues during genotyping process. To calculate heterozygosity we filtered out the HLA region (to avoid inflating the heterozygosity measured by linkage disequilibrium [LD]) in chromosome 6 and merged all chromosomes after selecting independent markers (pruning) with PLINK v1.9b3.32 (`--indep 50 5 2.5`).

Heterozygosity was calculated for each sample and any sample with values higher than 4 standard deviations (SD) from the mean heterozygosity were considered to be outliers. To avoid excluding individuals with inherent low heterozygosity as outliers, we also measured long runs of homozygosity (ROH), and considered as outliers only those with values below 4 SD of the residuals of the linear regression between heterozygosity and ROH. Heterozygosity and ROH were calculated with the PLINK commands `--het` and `--homozygous`, respectively.

We identified 199 samples as heterozygosity outliers (**Figure 4**). To further understand if these heterozygosity outliers were being driven by a higher missingness rate, we tested if heterozygosity was associated with missing rate levels. This appeared to be the case: samples with high missingness rates were also more heterozygous. Most of these samples were coming from three DNA plates (110, 111, and 279). The 199 heterozygous samples were excluded.

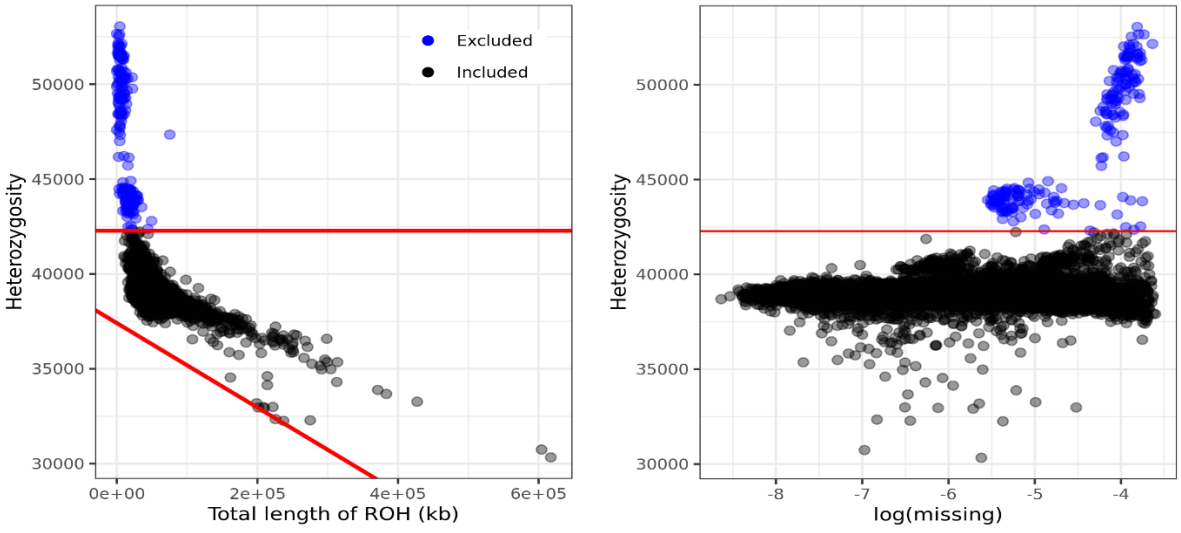

Figure 4. Heterozygosity depicted against runs of homozygosity (ROH) (left) and missingness rates (right). Red lines represent the filtering thresholds. Blue dots represent samples that are excluded based on more than 4 standard deviations (SD) above the mean heterozygosity or more than 4 SD of the residuals of the linear regression between heterozygosity and ROH below the predicted heterozygosity from this same linear regression analysis.

6. Sample mix-ups

Sample mix-up is investigated by looking at gender mismatch, where gender information of each sample as recorded in the Lifelines database is compared with genotypes at chromosomes X and Y. This method however does not detect same-sex sample mix-ups and is not reliable when there are sex chromosome abnormalities. Therefore we additionally used the familial relationships between Lifelines samples according to the Lifelines pedigree information and compared the expected genetic sharing with the genetic relationships of each pair of samples. Each potential sample mix-up detected was carefully analyzed and evaluated taking into consideration plate number and position as well as the supposed volunteer's questionnaire information regarding first- and second-degree relationship (children, partner, parents, and siblings) with other Lifelines members. The specific details on the gender mismatch and familial relationship concordance analyses are described below.

6a. Chromosome X QC and check

The markers on chromosome X were analyzed independently from the other chromosomes. We first extracted all samples that passed QC at this level of filtering (step 5). At the marker level, we first applied the same thresholds as for the autosomal chromosomes in steps 1 (i.e., removing duplicate markers [N=45, 0.2%]) and 2 (i.e., filtering by call rate [N=6,562, 29.3%]). Next we inferred genetically determined sample sex by calculating heterozygosity of chromosome X with PLINK (`--impute-sex`) using default thresholds (male: F>0.8, female: F<0.2). This result was later compared with respective sex information for each sample from baseline questionnaires. Samples with a mismatch between genetically determined sex and questionnaire sex information were flagged "Non-concordant", and samples that could not reach a sex definition from this calculation (i.e., 0.2<F<0.8) were flagged as "Failed sex imputation". Flagged samples were used together with the pedigree concordance analysis.

After full sex and familial information was ascertained we filtered chromosome X to remove markers with a MAF <0.02% (N=704, 4.5%) and HWE outliers ($p<1x10^{-10}$) with only females (N=912, 6.0%).

6b. Pedigree concordance analysis

The flow diagram of the pedigree concordance analysis is shown in **Figure 5**. For the pedigree concordance analysis the genetic autosomal data of the UGLI2 samples were merged with high quality genetic data of the CytoSNP and UGLI-GSA samples. Only markers with an imputation quality >0.95 were extracted from the available VCF files using BCFtools v1.16 and converted to PLINK binary format. Next the data of the three datasets (CytoSNP, UGLI-GSA, and UGLI2) were merged.

These data were then used to infer the relationship between each possible pair of samples using KING 2.2 (http://people.virginia.edu/~wc9c/KING/) with the commands `--relations --degree 2`. We compared this with the pedigree information available from the Lifelines database, which was optimized during sample selection. KING classifies the relationship between pairs as one in seven possibilities (Monozygotic twin / duplicates, Parent-offspring, Full siblings, 2nd degree, 3rd degree, 4th degree and Unrelated (sharing no genetic relationship)) according to the parameters of

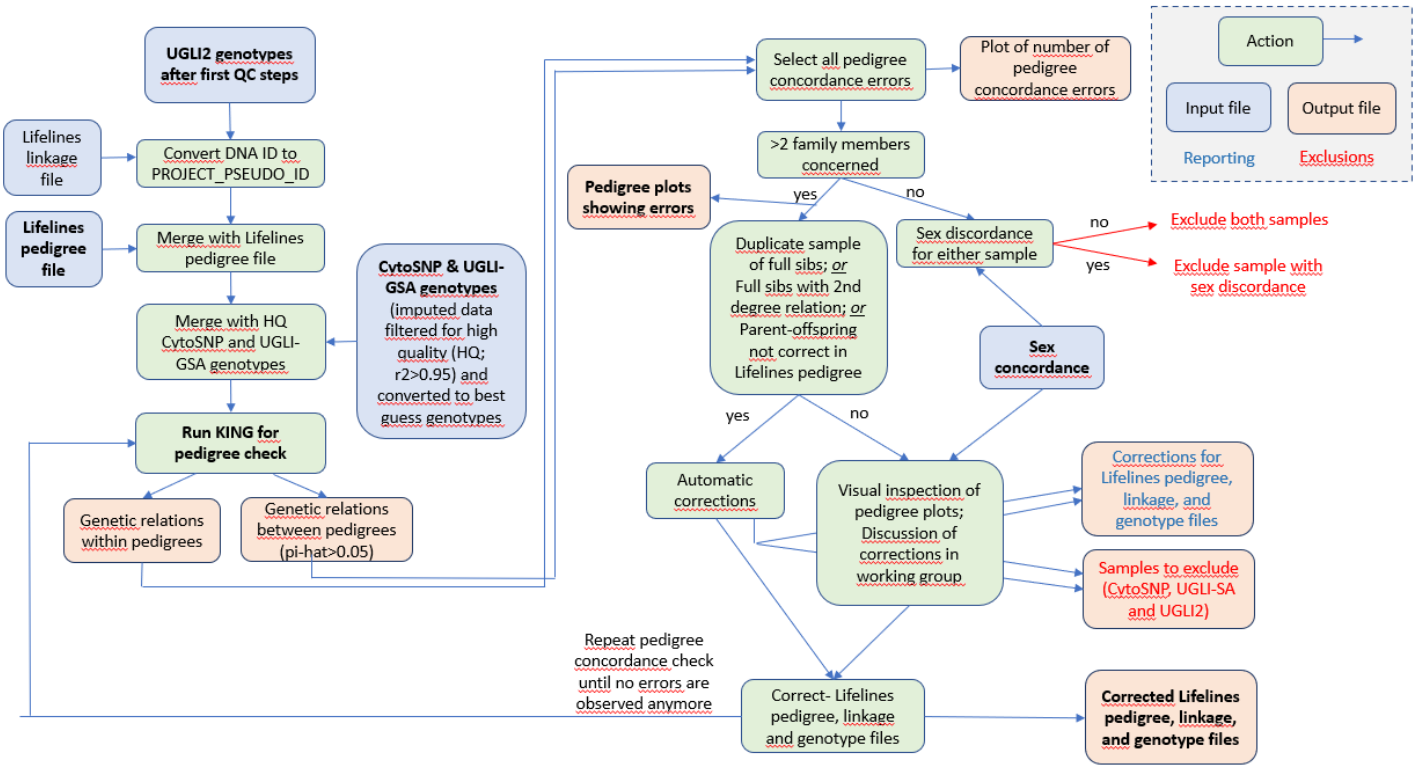

Figure 5: Flow diagram of the pedigree concordance analysis

genetic similarity described in Manichaikul *et al* (2010) (2). Additionally, it evaluates each of the relationships provided in the pedigree information, and flags each relationship not supported by genetic information.

In a first round of the pedigree concordance check, many errors in sex concordance and family relationships were observed for DNA plates DNA112, DNA113, and DNA114. Samples of plate DNA112 appeared to all be duplicate samples of those on plate DNA066. Samples of plate DNA113 were often found to be unrelated to family members, while samples of plate DNA114 were found to be related, and vice versa. We therefore decided to exclude all samples from plate DNA112 and swap the samples of plates DNA113 and DNA114. The number of sex mismatches and family errors decreased drastically after these decisions. Therefore we decided to exclude the samples of plate DNA112 and swap the samples of plates DNA113 and DNA114. The results presented above in steps 2-6a actually already concern this corrected dataset.

A total of 147,833 known family relationships were confirmed, while 18,689 (11.2%) relationships were flagged as errors (**Figure 6**). In addition 10,005 new relationships were found, of which 1,071 (10.7%) concerned first-degree relations (monozygotic twins/ duplicates, parent-offspring or full siblings). We analyzed any family relationship within families flagged as “error” that resulted in a genetically calculated first-degree or “unrelated” relationship (N=544, 0.3%), as well as the 1,071 first-degree relations between families. If an error occurred in a family with only two genotyped family members, the samples were checked for sex discordance and if there was a sex mismatch for one of the samples, this sample was excluded (N=36). In case of no sex discordance, both samples were excluded (N=104). For each of the families with errors and that had more than two genotyped individuals (N=1,239), we visualized the information in a pedigree plot, and we coupled it with the age, sex (according to pedigree and genetically determined), and questionnaire information on (pseudonymized) surnames, parental and offspring, and siblings’ birth dates, and parental death

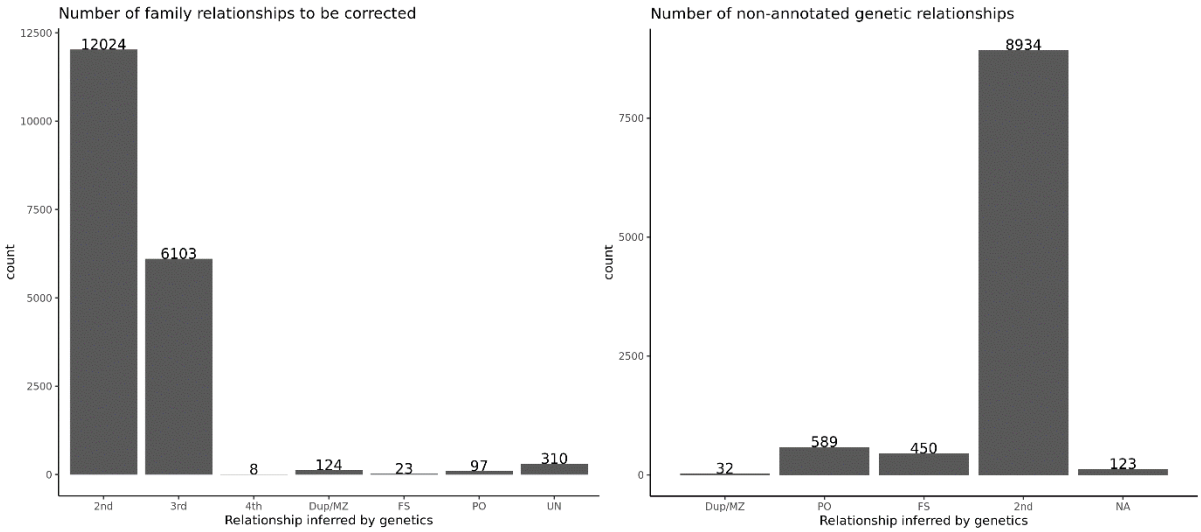

Figure 6. Summary of the genetic relationships calculation. Dup/MZ: duplicates /monozygotic twins, PO: parent-offspring, FS: full siblings, UN: unrelated, and ordinal numbers indicate relationship degree. Left: Family relationships flagged as errors, calculated genetic relationships are shown. Right: relationships not indicated by the family information and found with genetic calculation.

years (see example in **Figure 7**). An event indicated by the genetic relationship (be it error or new finding) was considered true, only if it was supported by the other independent layers of information, namely: 1) the same sample showed concordant genetic relationships across a family and/or in different generations, 2) age and sex (including sex-concordance, explained in the next section) made sense with the indicated familial relationship, or 3) the relationship was indicated directly or indirectly in the family information section of the questionnaire. If these layers reached to contradictory conclusions, the sample information would be changed according to the strongest evidence (i.e., if layer 1 applied but layers 2 and 3 did not, this could be considered a sample mix-up). Each event was looked carefully and all the decisions and evidences are reported in detail.

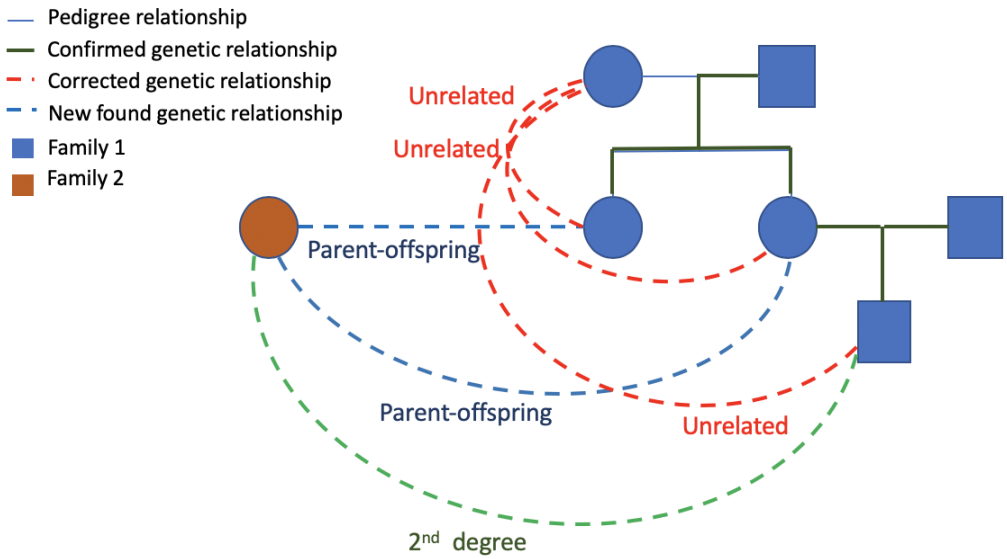

Figure 7. Example pedigree analysis of a sample mix-up. The participant from family 2 (in gold) is actually the grandmother of family 1, while the supposed grandmother of family 1 (in blue) does not belong to this family.

The pedigree concordance analysis revealed that 144 errors occurred due to real monozygotic twins; 178 were full sibs that genetically turned out to be half sibs; 84 were corrections of parents within a family (i.e. a dummy was assigned, but the actual parent was present or full sibs that were thought to be half sibs); and 135 were identified as sample mix-ups. Of these sample mix-ups, 66 had enough genetic sharing with other Lifelines volunteers (i.e., relationships) to be reliably assigned to the correct individual (and family). The rest of the mixed-up samples (N=69) were excluded. Lists of samples swaps and samples to be excluded were created and with these files the Lifelines pedigree file, linkage files, and genotype PLINK fam files were corrected. The pedigree concordance analysis was repeated using these new files and verified that no additional sample mix-ups were present after this correction process. During the process we decided not (yet) to merge groups of families in which no other errors than duplicate samples between families occurred, since this would only affect the Lifelines pedigree file and not whether UGLI2 samples should be swapped or excluded. Therefore there are still 308 groups of 2-6 families with first-degree relations that could be merged.

After this step we removed in total 156 samples that failed the pedigree concordance check as well as 97 samples still flagged as "Non-concordant" by sex, leaving 28,249 samples for the population stratification analysis.

7. Population Stratification

Population stratification of the UGLI2 cohort was performed in similar fashion to population stratification by the UK-Biobank on all autosomes of the UGLI2 samples, using PLINK, GCTA (https://yanglab.westlake.edu.cn/software/gcta/), and the populations as defined by the 1000-genomes (1000G) cohort (https://www.internationalgenome.org/). All variants with a minor allele frequency (MAF) < 0.01 were excluded for this analysis.

Next, high LD regions as defined by the UK-biobank were removed, and only bi-allelic SNPs with single-nucleotide alleles were retained. Because UGLI2 genotyping data were generated in human genome build hg38, 1000G data was lifted over from hg19 to hg38 using UCSC's liftOver tool (https://genome.sph.umich.edu/wiki/LiftOver). Variants mapping to the sex chromosomes or without new coordinates were removed from further analysis. Variant IDs were matched between the UGLI2 and 1000G cohorts based on chromosome and position. All variants with duplicate IDs and non-matching alleles were removed after selection of common variants in both cohorts. 1000G variants were pruned using PLINK (*--indep-pairwise 1000 5 0.2*), and both cohorts were filtered to only keep a final selection of 194,491 common and pruned variants.

A genetic relationship matrix was created for the 1000G cohort (without the admixed AMR population samples) and used for principle-component analysis (PCA) of up to 20 principle components (PCs) to generate PC-loadings that were projected onto the UGLI2 cohort. By examining up to 20 PC eigenvalues and their individual contribution to outlier detection we decided on a cut-off of 4 standard deviations from the centroid of each of the first five PCs (**Figure 8**), resulting in identification of 142 samples as genetically non-European.

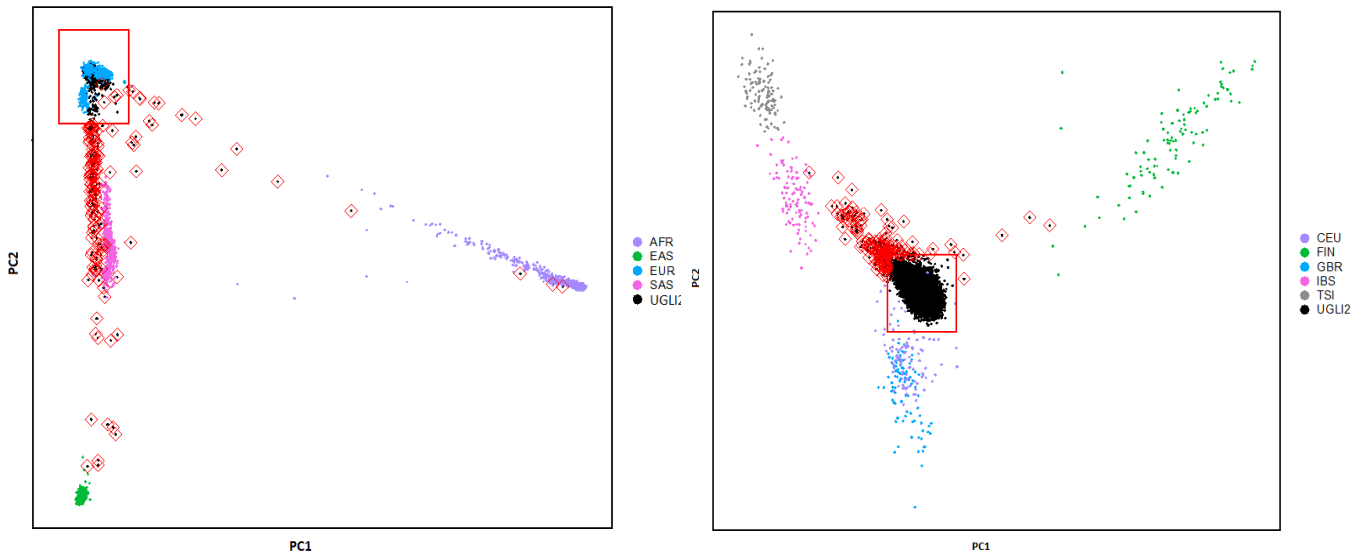

Figure 8: Population stratification analysis of UGLI2 samples. In the left plot, the principal component (PC) analysis of all 1000G superpopulations (left) identified 142 non-Europeans in UGLI2. A non-European was defined as >4 SDs from the centroid of the 1000G European population (blue dots) for the first five PCs. In the right plot, the PC analysis using only 1000G European populations identified 161 genetic outliers within the UGLI2 cohort. An UGLI2 genetic outlier was defined as >4 SDs from the centroid of all UGLI2 samples (blue dots) for the first two PCs. The 4SD boundaries are marked by the red boxes. All non-European UGLI2 samples or UGLI2 genetic outliers are marked by red diamonds. Note that the left figure only shows PC1 and PC2, while five PCs were used for identification of non-European samples.

To assess the population structure of the UGLI2 cohort within the European population we re-created a GRM, PCA of up to 20 PCs, and PC-loadings of only the 1000G European population (503 samples) as classified by 1000G. A total of six variants that are mono-allelic in the 1000G European population had to be removed from the PC-loadings before final projection onto the UGLI2 cohort. The previously described stratification method was applied, 4SDs from the centroid of the UGLI2 cohort in the first two PCs resulted in identification of 161 UGLI2 outliers within the European population according to two PCs (**Figure 8**).

Non-European UGLI2 samples and UGLI2 genetic outliers were not removed from the dataset, but lists are available in the files '*nonEuropeans.flagged.samples*' and '*UGLI2_genetic_outliers.flagged.samples*', respectively. It is up to the researcher if he/she wants to remove them or correct for population stratification in his/her genetic analysis.

8. Mendelian errors

After establishing the family relations within the combined set of CytoSNP, UGLI-GSA and UGLI2 samples, we quantified the number of mendelian errors detected per SNP. A Mendelian error is a discrepancy between the genotypes observed in parents and their offspring. For example, for SNP x, both parents have an AA genotype, however their children report a BB or AB genotype. This discrepancy would be flagged as a Mendel error, as children cannot have inherited allele B from

their parents. We identified Mendel errors using PLINK and the *--mendel* command, and then counted how many errors were observed for each SNP. No SNPs with more than 1% of Mendelian errors across all Parent-Offspring (PO) pairs were observed and hence no SNPs were at this step.

9. Hardy-Weinberg equilibrium in unrelated individuals

Lastly, we re-calculated HWE p-values per SNP including only unrelated individuals within UGLI2. To generate the subset of unrelated individuals, first the data was LD pruned using PLINK (*--indep-pairwise 1000 5 0.1*). We next used GCTA to calculate the genetic relationship matrix and let GCTA decide on the optimal subset of individuals such that there were no first- and second degree relatives within this subset (pi-hat<0.15). To determine the HWE p-values, we used PLINK and the command *--hardy*, same as in steps 4 (autosomal markers) and 6a (X chromosomal markers) in this QC protocol, but now on the subset of unrelated individuals for the autosomal markers and the females among the subset of unrelated individuals for the markers on the X chromosome, respectively. All genetic markers with a HWE p-value ≤ $1x10^{-6}$ were excluded (N=960 autosomal markers and N=486 X chromosomal markers) leaving 502,257 and 13,113 markers on the autosomal and X chromosome, respectively.

10. Batch differences

The genotypes of the UGLI2 samples were called in 12 batches of 25 plates. We compared the QC results visually between the individual batches and overall found no significative differences between the batches.

Some slight differences between plates were observed in the percentages of samples excluded based on the stringent call rate threshold (with high percentages for plates 110, 111 and 269), but these didn't seem to be attributable to the batch. The heterozygosity rates were slightly higher for plates 110, 111 and 116.

11. Alignment with HRC

As a pre-imputation step the genetic markers were aligned with those available in the Haplotype Reference Consortium (HRC) dataset version v1.1 (http://www.haplotype-reference-consortium.org/site) using the tool '*HRC-1000G-check-bim-NoReadKey2.pl*' version 4.2.13 (McCarthy Tools (ox.ac.uk)). To use this tool first the positions of the genetic markers in the UGLI2 dataset were lifted over to genome build GRCh37.

The tool checks each marker for strand, alleles, position, reference and alternative allele assignments, and MAF differences. For the latter check allele frequencies were calculated on the final UGLI2 dataset. The tool produces files for each of these steps in order to (i) exclude unmapped markers (which include insertion/deletion polymorphisms); (ii) exclude SNPs with differing alleles; (iii) exclude palindromic markers with a MAF > 40%; (iv) update alleles to align with the positive strand; (v) update position; and (vi) update reference and alternative alleles to match those on the

imputation server. We decided not to exclude markers that had a MAF difference with the HRC dataset.

With this step 52,337 genetic markers (24,593 indels; 25,738 unmapped; 1,250 palindromic; 956 non-matching alleles[1]) were removed prior to imputation, leaving 450,110 autosomal markers and 12,621 X chromosomal markers in the final dataset.

12. Genetic imputation

A final set of 28,250 samples and 462,731 markers on autosomal and X chromosomes passing all QC steps described above were used for genetic imputation. Genetic imputation was done through the Sanger imputation service using the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org) panel.

[1] Overlap between the markers

Summarizing table

QC step	Number of variants						Number of samples			
	autosomes + XY			*chr X*						
	remaining	**excluded**	**%**	**remaining**	**excluded**	**%**	**remaining**	**excluded**	**%**	**flagged**
Pre-QC (incl Y and MT)	620834	-		22405	-		29166	-		-
Removed plate DNA112	620834	0	0,0%	22405	0	0,0%	29073	93	0,3%	
Exclude chr Y and MT	619715	1119	0,2%	22405	0	0,0%	29073	0	0,0%	-
Duplicate markers & samples	615210	4505	0,7%	22384	21	0,1%	28875	198	0,7%	-
Callrate <80%	615038	172	0,0%	22155	229	1,0%	28874	1	0,0%	-
Callrate <highcr*	593575	21463	3,5%	15798	6357	28,4%	28701	173	0,6%	-
MAF < 0.02%	510380	83195	14,0%	15094	704	4,5%	28701	0	0,0%	-
HWE p < 1E-10	503217	7163	1,4%	14197	897	5,9%	28701	0	0,0%	-
Sample heterozygosity	503217	0	0,0%	14197	0	0,0%	28502	199	0,7%	-
Relatedness check	503217	0	0,0%	14197	0	0,0%	28346	156	0,5%	-
Sex check	503217	0	0,0%	14197	0	0,0%	28250	96	0,3%	-
PCA analysis	503217	0	0,0%	14197	0	0,0%	28250	0	0,0%	142
Mendelian errors	503217	0	0,0%	13599	598	4,2%	28250	0	0,0%	-
HWE p < 1E-6 in unrelateds	502257	960	0,2%	13113	486	3,6%	28250	0	0,0%	-
Alignment HRC	450110	52147	10,4%	12621	492	3,8%	28250	0	0,0%	-