UGLI2+3 Quality Control Report (release 1.0 – Oct 2025)

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 second release of UGLI2 (officially UGLI2+3) comprising the genotype of 63,553 participants assessed using the FinnGen Thermo Fisher Axiom® custom array. In this QC screening we included all genotyped samples, and we focused on QC of genetic markers on the autosomes and chromosomes X (N=615,682 and 22,346 markers, respectively).

In brief, first sample specific priors for the genotype calling algorithm were generated using the *simple_ssp* tool provided by Thermo Fisher. Next, the genotypes were called using the Axiom Analysis Power Tools (APT) developed by Thermo Fisher and converted to binary PLINK format to perform the QC. The QC steps 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 using the two previous genotyping chips (CytoSNP 250k and the Infinium Global Screening Array® (GSA) MultiEthnic Disease Version) were used. Next the data were investigated for batch effects. Samples and variants that showed differences between genotyping batches were removed. Subsequently, we ascertained Mendelian errors and further removed genetic markers that deviated from HWE in unrelated individuals. Finally, population stratification was inspected by a principal component 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. Variant calling

Genotype calls for the autosomal, pseudo-autosomal chromosome XY, non- pseudoautosomal regions of chromosome X, chromosome Y and mitochondrial (MT) genetic variants were generated from Affymetrix CEL files using Thermofisher's Axiom Analysis Power Tools (APT). The genotypes for UGLI2 and UGLI3 were separately called using the exact same pipeline in batches of 50, 25, or 12 plates (n=952-4800 samples). The following genotyping settings were applied using Thermofisher's APT software: 1) The Dish Quality Control (DQC) cut-off remained unchanged at >0.82 for sample inclusion (412 samples were excluded), 2) The QC-call-rate cut-off was lowered from >0.97 to >0.90 after observing only a minimal drop in variant-calling quality while including many more samples (314 samples were excluded), 3) The whole-plate exclusion cut-off of >0.985 average QC-call-rate per plate was entirely removed. Finally, a sample specific priors file was generated by Thermo Fisher's *simple ssp* tool using the first 25 UGLI2 plates that seemed to have performed well, and supplied for the genotype calling algorithm during the variant calling. All other settings were applied according to the manufacturer's manual.

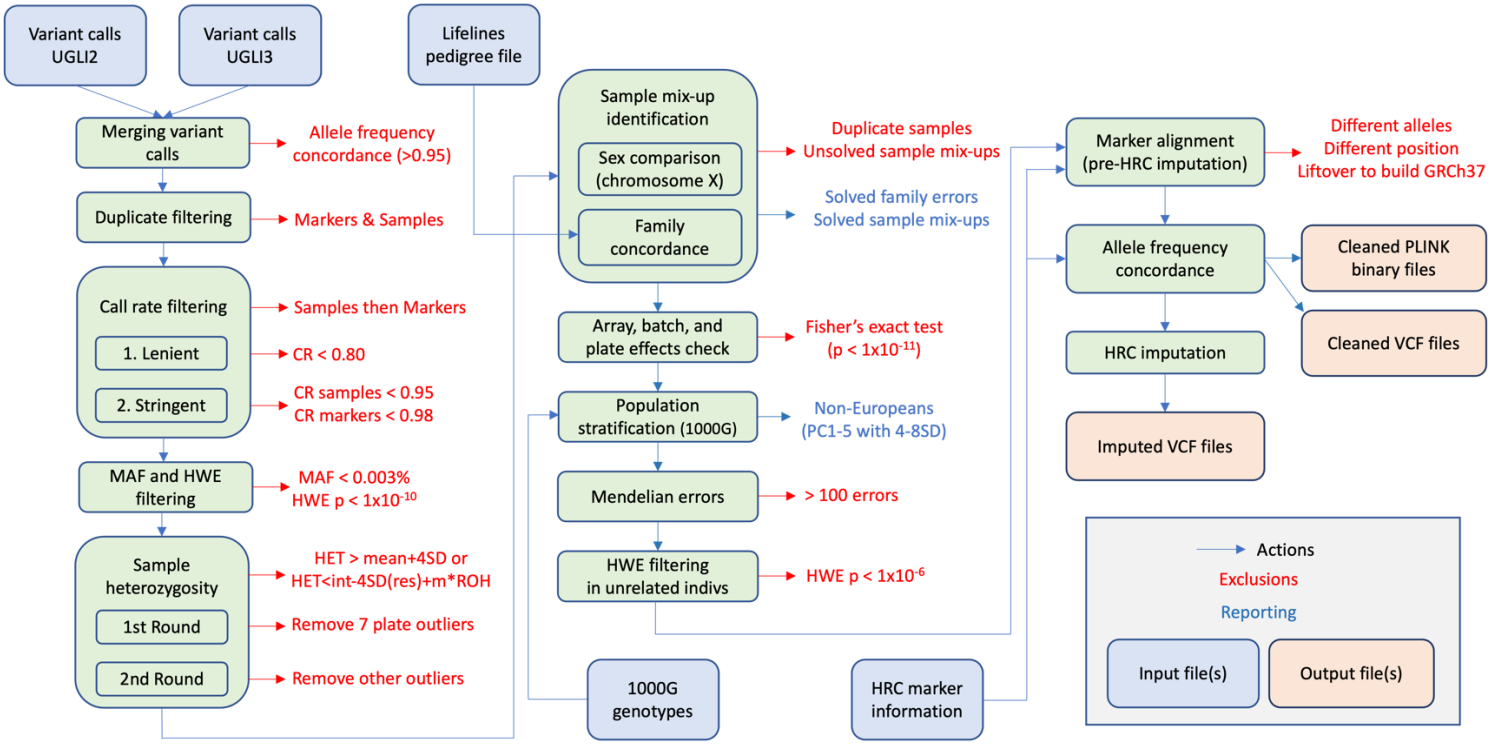

Figure 1. Steps and metrics evaluated in the quality control of the UGLI2+3 genotype data.

2. Merging variant calls

Assigned genotypes were converted to PLINK binary files (https://www.cog-genomics.org/plink/1.9/). Batches were merged per UGLI2 or UGLI3 array and separated into chromosomes (autosomal 1-22, X, Y, XY, and MT) to be further processed. Next, per chromosome the UGLI2 and UGLI3 variant calls were merged. Variants with missing genotypes were excluded (107 autosomal variants) and allele frequency concordance between UGLI2 and UGLI3 variant calls were evaluated (**Figure 2**). Variants with more than 5% discordance in allele frequencies between the UGLI2 and UGLI3 variant calls were flagged (965 autosomal variants, 29 variants on chromosome X). These flagged variants and samples were evaluated for removal by the end of the QC pipeline (but in fact all appeared to already have been removed during other QC steps).

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 445 mitochondrial markers were excluded.

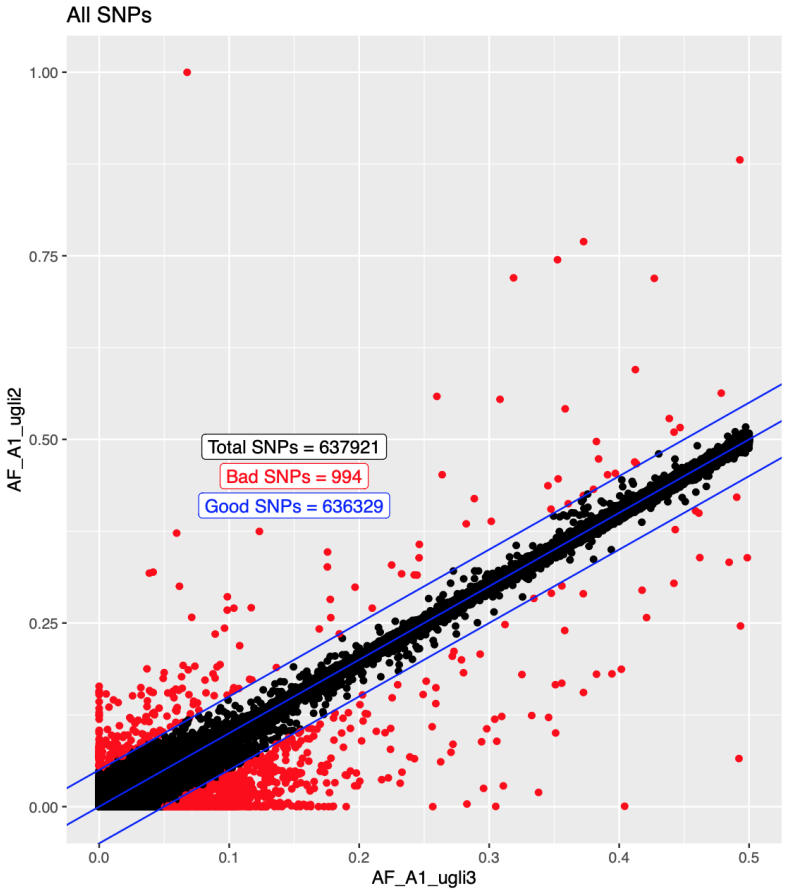

Figure 2. Allele frequency (AF) concordance test between the UGLI2 and UGLI3 called genotypes. We test for a 5% deviation in AF in either direction (blue lines), the black dots are variants within the accepted AF concordance and the red variants are flagged as outliers.

3. Filtering duplicate markers and samples

Duplicate markers and samples were removed for autosomal and pseudo-autosomal chromosomes. To do this, marker names were converted to chr_pos_A1_A2 ids, where A1 and A2 are the two alleles in alphabetic order. This way, tri-allelic markers are identifiable through two different markers, as are single nucleotide polymorphisms (SNPs) and insertion-deletion polymorphisms (indels) whose positions overlap. Upon identifying duplicate markers they were assigned 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 calculated with the *--missing* command. Next, the marker with the 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 samples known to be duplicates as derived from their sample IDs 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 983 and 9 duplicated (by position and allele) autosomal and chromosome X markers, respectively, and 286 duplicated samples.

4. Filtering markers and samples with a low call rate

Autosomal and pseudo-autosomal markers with high missing rates 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 (2% for markers and 5% for samples). We set the threshold for samples slightly lower than before to retain as many samples as possible, assuming that the missing genotypes can be imputed with sufficient quality. All the steps here were done using the *--missing -- remove* and *--exclude* PLINK commands, following this workflow: 1) Calculate the missing rate per sample and remove samples with a missing rate >20%; 2) Calculate the missing rate for markers and remove markers with a missing rate >20%; 3) Recalculate the missing rate for samples and remove samples with a missing rate >3%; 4) Recalculate the missing rate for markers and remove markers with a missing rate >1%.

The lenient call rate filter (80%, i.e. missing rate=20%) excluded 66 autosomal and 1 chromosome X markers and no samples. The stringent call rate filter excluded 12,829 autosomal and 281 chromosome X markers and 429 samples.

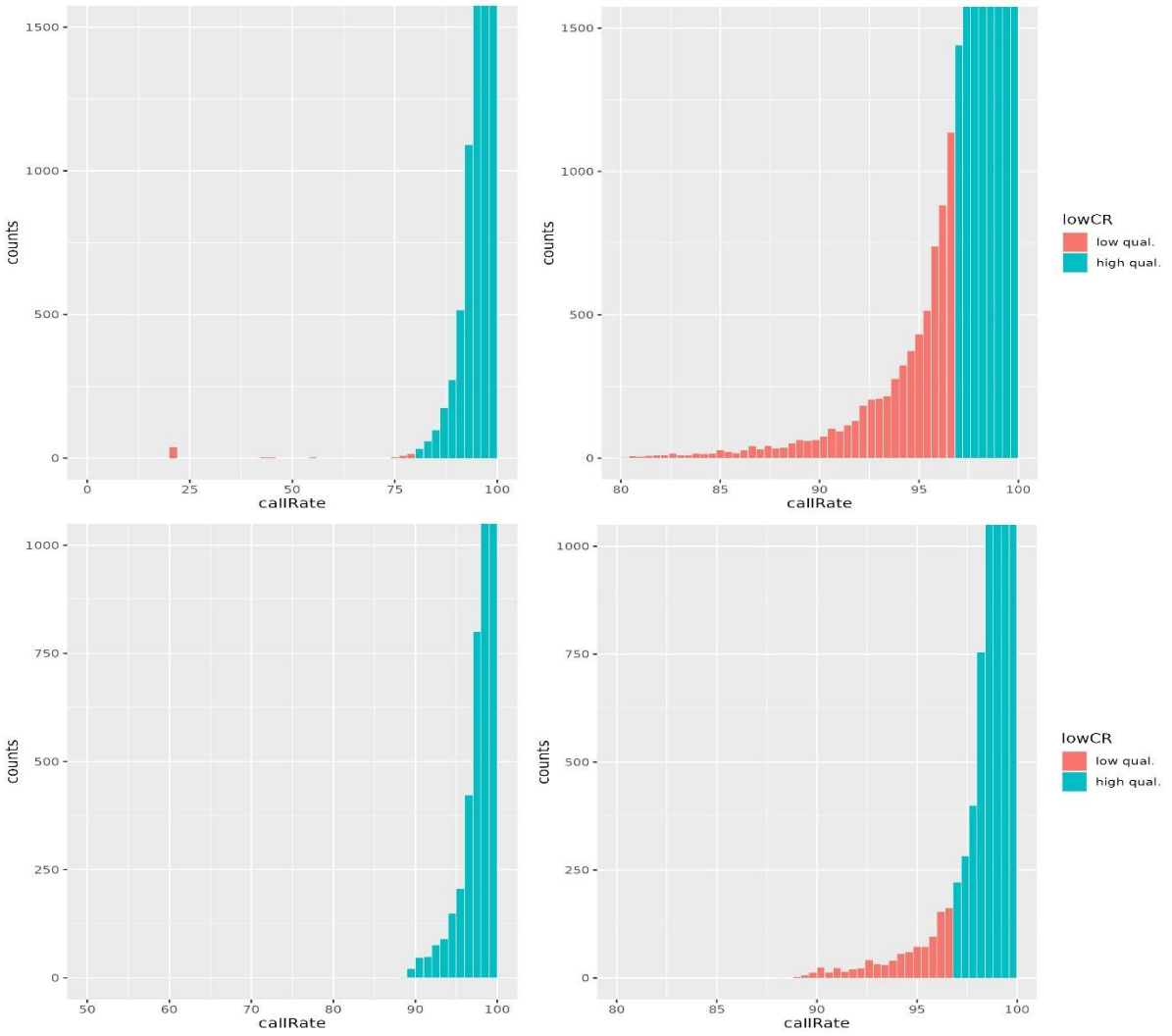

Figure 3: Distribution of the call rates before (on the left) and after (on the right) lenient filtering. The top graphs represents the marker call rates; the bottom ones the sample call rates. Markers or samples below the respective threshold are shown in red, those above in blue.

5. Filtering 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 MAF <0.003% (three or less minor alleles) and/or markers with a HWE p-value < $1x10^{-10}$ were discarded because they were considered uninformative and of poor quality. 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 11).

A total of 55,695 (9.27%) and 1,622 (7.94%) autosomal and chromosome X markers, respectively, were found to have a MAF below the threshold, and 16,249 (2.98%) autosomal and 185 (0.91%) chromosome X markers were out of HWE.

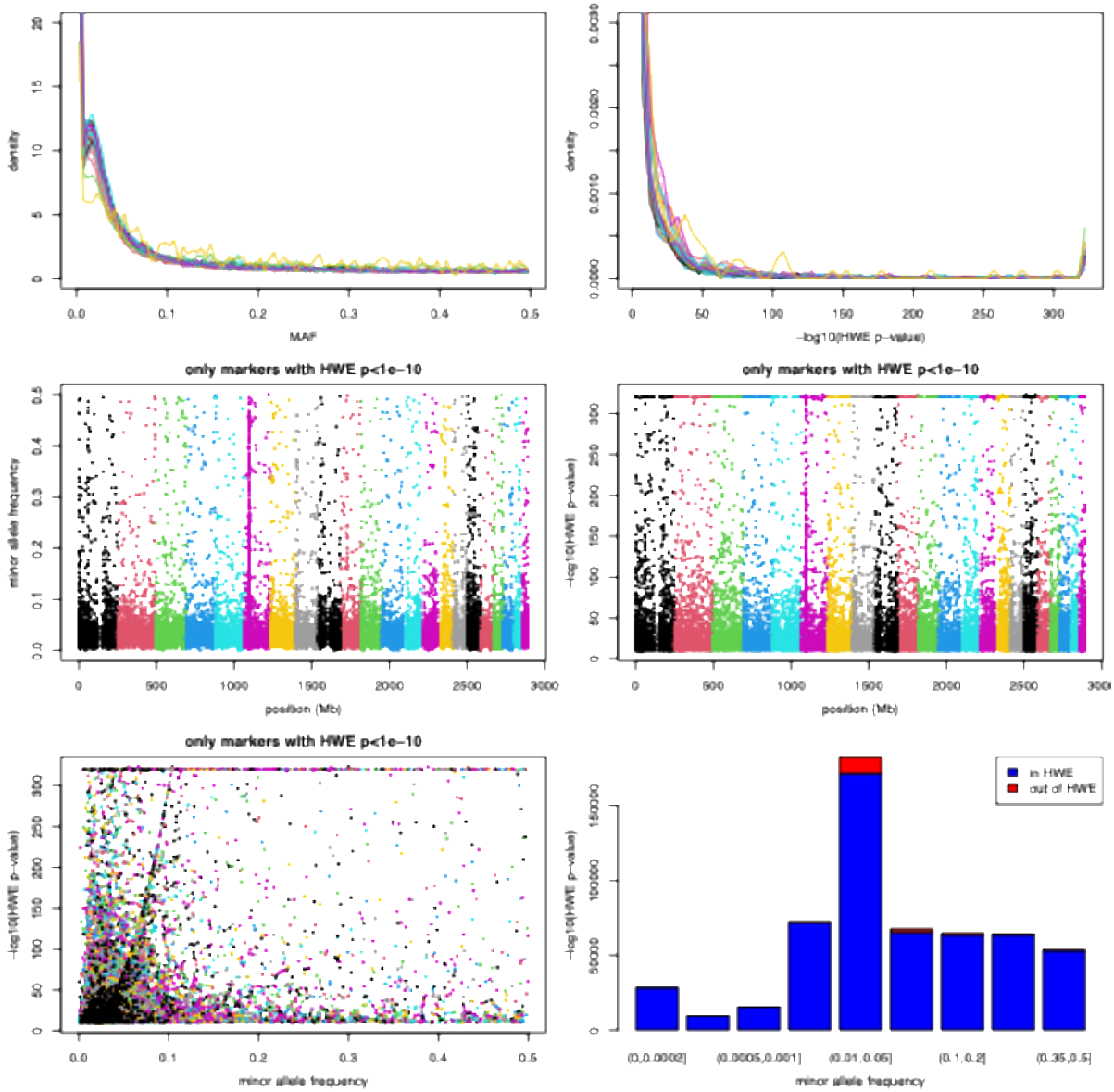

Figure 4: Various graphs showing the distribution of and relation between minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) p-value per chromosome. Upper left and upper right plot show the distributions of the MAF and HWE p-value, respectively. The middle plots show the MAF (left) and -log_{10}(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.

6. Sample heterozygosity

A common step in quality control of genome-wide arrays is to check for sample heterozygosity. Outliers showing excess or depletion in heterozygous genotypes may be due to DNA contamination or issues during the 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 a large number of samples as heterozygosity outliers, including many extreme outliers that appeared to negatively affect the SD cut-off we set, preventing us to reliably QC on heterozygosity (**Figure 5, top graphs**). Our more lenient variant calling cut-offs (see step 1) have included many additional samples of lower quality, including samples of too low quality that we identify here. Upon assessing the heterozygozity per plate we identified seven plates with a large number (or entire plates) of extreme outlier samples. Assuming these seven plates were structurally exposed to DNA contamination or genotyping errors, we decided to exclude them entirely from further QC (408 samples, DNA plates 110, 111, 377, 378, 379, 428, and 527). Next, we performed a second round of heterozygosity and ROH QC step as described above to exclude moderate outliers, in this step 858 outlier samples (1.39%) were excluded (**Figure 5, bottom graphs**).

7. 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 relationships (children, partner, parents, and siblings) with other Lifelines members. The specific details on the gender mismatch and familial relationship concordance analyses are described below.

7a. 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 6). At the marker level, we first applied the same thresholds as for the autosomal chromosomes in step 1 (i.e., removing duplicate markers [N=9, 0.04%]) and step 2 (i.e., filtering by call rate [N=282, 1.28%]). 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.

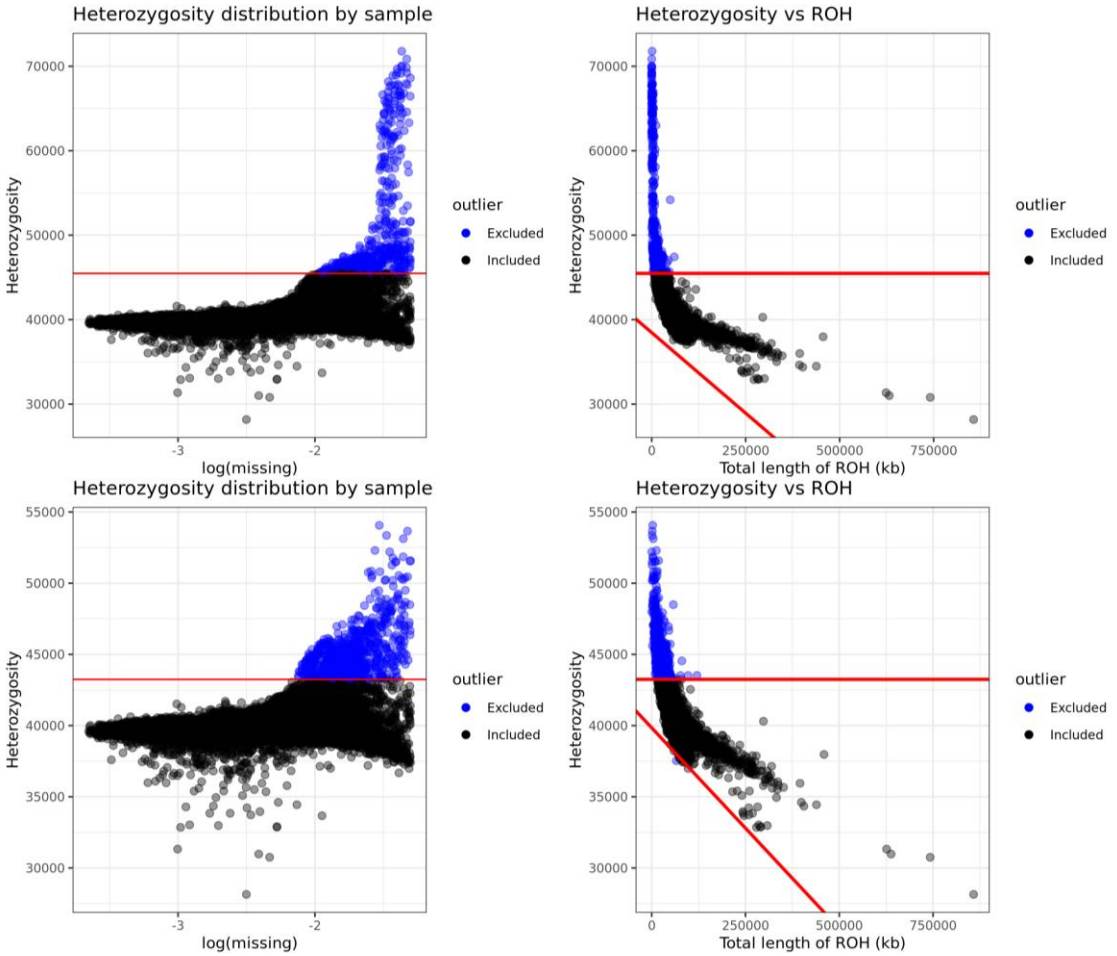

Figure 5. The top graphs depict round 1, and the bottom graphs round 2. Heterozygosity depicted against missingness rates (left) and runs of homozygosity (ROH) (right). Red lines represent the filtering thresholds. Blue dots represent samples that are excluded based on more than 4 standard deviations (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.

After full sex and familial information was ascertained, we filtered chromosome X to remove markers with a MAF <0.003% (less than three minor alleles, N=1622, 7.94%) and HWE outliers ($p<1x10^{-10}$) with only females (N=185, 0.91%).

7b. Pedigree (family) concordance analysis

The flow diagram of the pedigree concordance analysis is shown in **Figure 6**. For the pedigree concordance analysis, the genetic autosomal data of the UGLI2+3 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+3) 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)). Additionally, it evaluates each of the

relationships provided in the pedigree information, and flags each relationship not supported by genetic information.

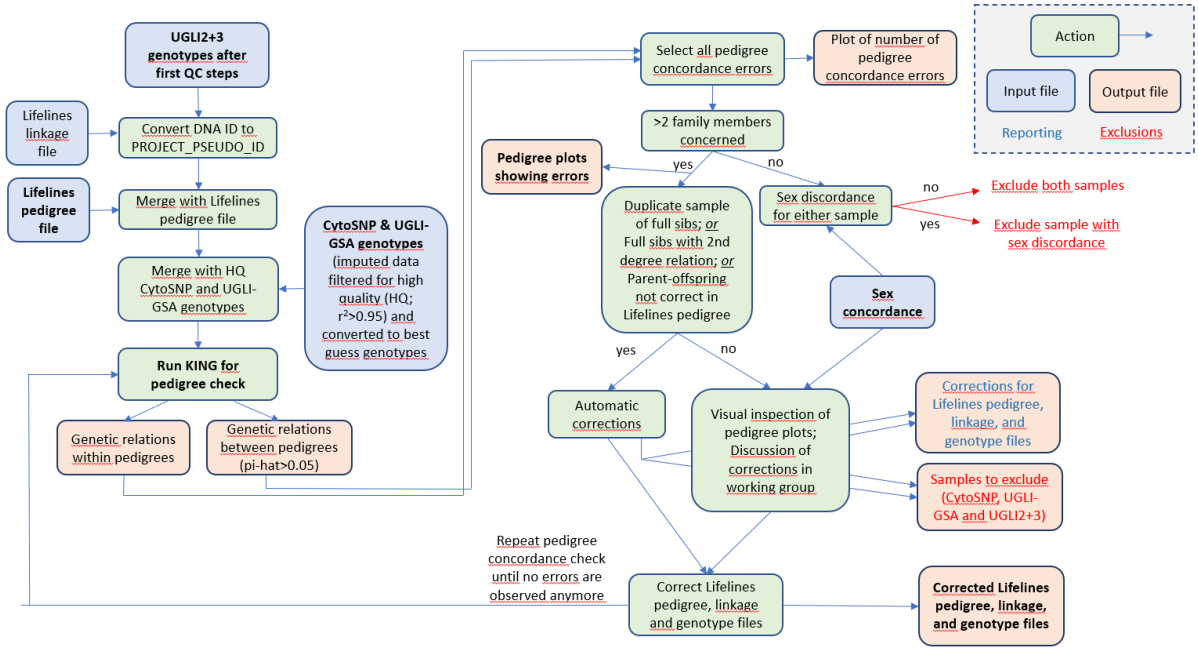

Figure 6: Flow diagram of the pedigree concordance analysis

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 to these family members, 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 keep this change. The results presented above in steps 2-6a actually already concern this corrected dataset.

A total of 232,612 known family relationships were confirmed, while 1,252 (0.54%) relationships were flagged as errors (**Figure 7**). In addition, 20,611 new relationships were found, of which 2,690 (13.1%) concerned first-degree relations (monozygotic twins/ duplicates, parent-offspring or full siblings). We analyzed any family relationship within families flagged as “error” that concerned a genetically calculated first-degree or “unrelated” relationship (N=348+509), as well as the 2,690 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=56). For each of the families with errors and that had more than two genotyped individuals (N=302), 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 years (see example in **Figure 8**). 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.

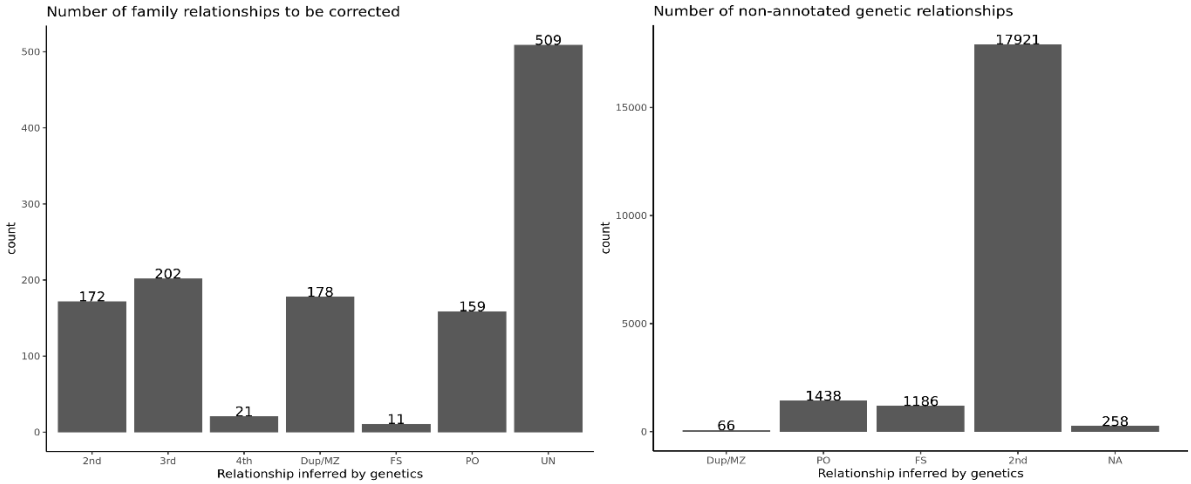

Figure 7. Summary of the genetic relationships calculation. Dup/MZ: duplicates /monozygotic twins, PO: parent-offspring, FS: full siblings, UN: unrelated, and ordinal numbers indicate relationship degree. Left: Family relationships flagged as errors, calculated genetic relationships are shown. Right: relationships not indicated by the family information but found with KING.

The pedigree concordance analysis revealed that 186 errors occurred due to real monozygotic twins; 40 were full sibs that genetically turned out to be half sibs; 11 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 288 were identified as sample mix-ups. Of these sample mix-ups, 216 had first-degree 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=62) 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+3 samples should be swapped or excluded. Therefore, there are still 794 groups of 2-6 families with first-degree relations that could be merged.

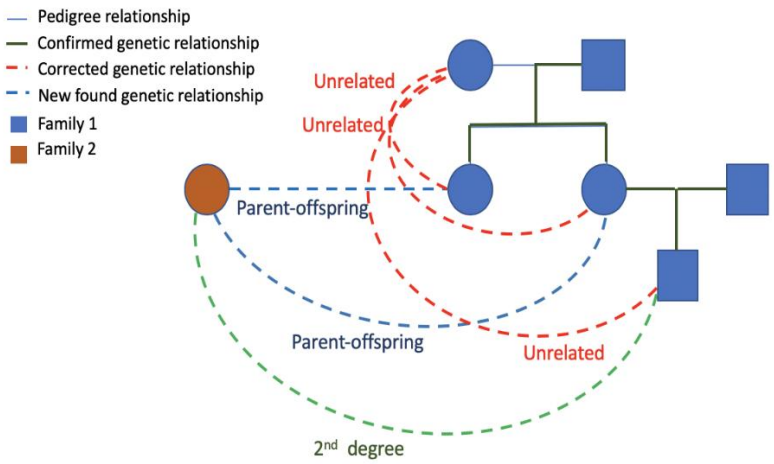

Figure 8. Example pedigree analysis of a sample mix-up. The participant from family 2 (in gold) is actually the grandmother of family 1, while the supposed grandmother of family 1 (in blue) does not belong to this family.

After this step we removed in total 469 samples (0.77%) that failed the pedigree concordance check as well as 220 samples (0.36%) still flagged as "Non-concordant" by sex, leaving 60,157 samples for the population stratification analysis.

8. Array, batch, and plate effects

We simultaneously performed three tests, UGLI2 versus UGLI3, batch (of 12-50 plates), and plate effects, to control for consistency across experimental setup in similar fashion to the UK Biobank (UKB) (Bycroft et al., Nature, 2018). In short, within the same population we would not expect genotype frequency differences, these tests are performed to identify markers with unusual genotype frequency differences between UGLI2 and 3, between batches and plates. Genotyping inaccuracies, sample quality, or other technical problems can be reasons why unusual genotype frequencies may be observed. The UGLI2 and UGLI3 samples were genotyped on the same genotyping array and the same sequencing equipment but were processed at different moments in time and by different technicians, hence we decided to test for UGLI batch effects.

Similar to the approach employed by the UKB, we used a Fisher's exact test on the 2x3 table of genotype counts (or 2x2 table for haploid markers). Markers that failed the UGLI batch effects test were excluded from the full dataset; markers that failed the batch or plate effects tests were set to missing for only the samples in failed batch or plate concerned.

8.1 P-value threshold

The same approach as the UKB was applied to determine an associated p-value for all three (UGLI batch, batch, and plate) hypothesis tests. We used a p-value threshold of 10^{-11}, any marker with a smaller p-value is considered as failing the test. The UGLI2+3 dataset contains 36 batches, 19 plates per batch on average, and ~638,000 markers, making a total of around 4.4×10^{-8} tests. We multiplied this by the UKB assumed family-wise error rate of 0.005 and rounded it up to a p-value threshold of 10^{-11}, because Lifelines has a stronger family structure than the UKB. As stated in the UKB publication, this threshold allows us to only exclude a marker or set it to missing if there is strong evidence for deviation from the null hypothesis.

8.2 Chromosome X markers

We tested both the haploid and diploid chromosome X markers in three-fold to account for possible gender imbalances affecting some marker's p-value, again according to the UKB approach. We ran each test using 1) female samples only (diploid), 2) male samples only (haploid), and 3) both genders combined (all samples), as inferred by the Affymetrix APT software based on each sample's genotype. The smallest p-value out of all three subsets was compared to the p-value threshold for marker exclusion or to set markers as missing. We excluded the combined sample results of chromosome X for the UGLI batch effects test, because p-values were strongly affected by the imbalance in male/female samples.

8.3 Exclusions and additional call rate filtering

A total of 24,016 and 432 autosomal and chromosome X markers, respectively, were excluded by the array effects test (**Figure 9**) and 38,560 and 1,119 autosomal and chromosome X markers, respectively, were set to missing for the samples of one or more batches or plates. This resulted in a low call rate for certain markers, hence, an additional 2,563 and 49 autosomal and chromosome X markers, respectively, were removed from the full dataset because their call rate was below 98%.

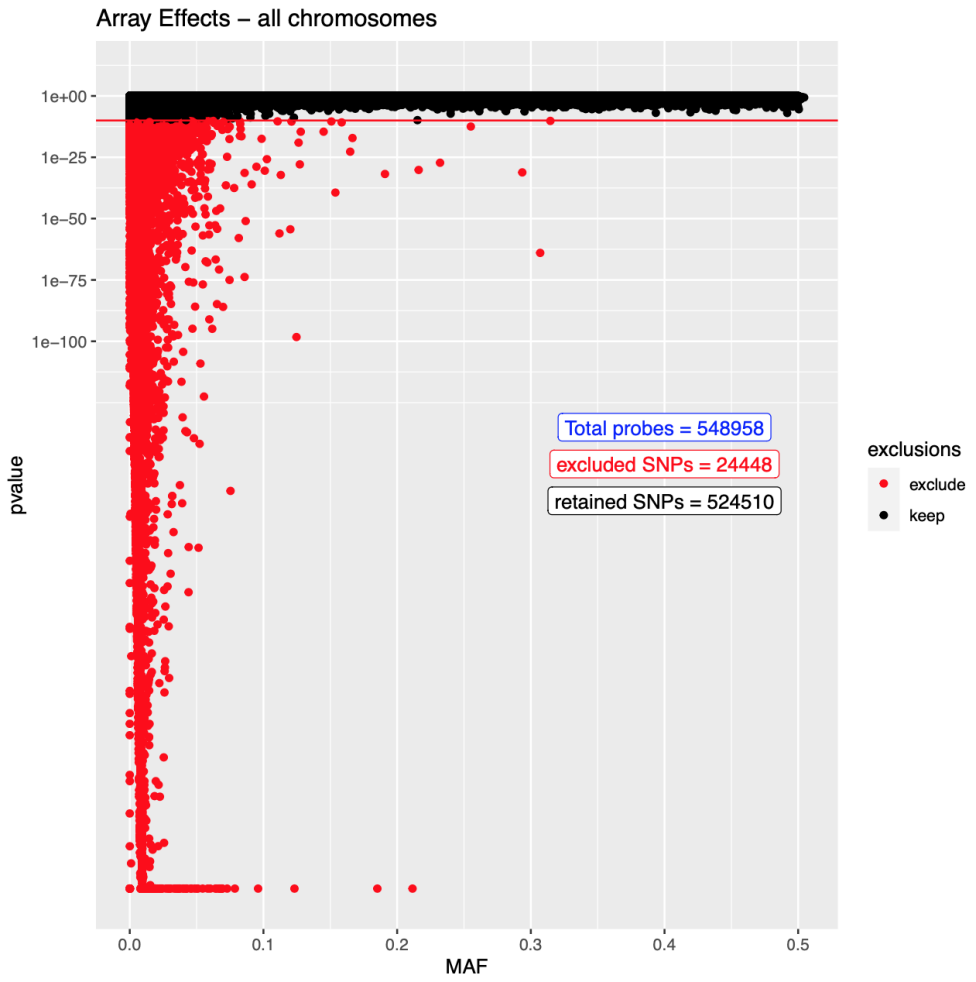

Figure 9. Scatterplot of the array effects test by plotting the Fisher's exact test's p-value (y-axis) vs the MAF (x-axis) of each marker. All markers that failed the test are marked in red.

9. Population Stratification

Population stratification of the UGLI2+3 cohort was performed in similar fashion to population stratification by the UKB (Bycroft et al., Nature, 2018). In short, we used the analysis tools PLINK and GCTA (https://yanglab.westlake.edu.cn/software/gcta/) to generate a genetic relationship matrix (GRM) for the 1000-genomes (1000G) cohort (https://www.internationalgenome.org/). Next we performed a principle component analysis (PCA) of the 1000G cohort to generate PC-loadings of up to 20 PCs upon which the UGLI2+3 cohort was projected to define global and within-Europe populations.

We included only the autosomal variants in this analysis, all variants with a MAF < 0.01 were excluded, and only bi-allelic SNPs with single-nucleotide alleles were retained. High linkage disequilibrium (LD) regions (long-range LD blocks), as defined by the UKB, were removed. The 1000G data was lifted-over from hg19 to hg38 using UCSC's liftOver tool (https://genome.sph.umich.edu/wiki/LiftOver), because the UGLI2+3 genotypes were generated in human genome build hg38. We pruned the 1000G variants using PLINK (*--indep-pairwise 1000 5 0.2*) and performed the final analysis on the set of pruned and common SNPs between 1000G and UGLI2+3.

By examining up to 20 PC eigenvalues and their individual contribution to outlier detection we decided to only use the first five PCs and to apply a dynamic cut-off of 4-8 standard deviations (SD) from the centroid of each PC (**Figure 10**). A single SD cut-off does not account for variation caused by some

(global) populations clustering more densely than others, hence we used a dynamic cut-off per population based on visual representation.

9.1 Global populations

We excluded the AMR population from our analysis because of the heterogeneity of this population and the difficulty to accurately classify individuals in this population. In total, we identified 300 individuals as genetically non-European (**Figure 10, left-side**). The samples concerned were not removed from the dataset but instead are listed in the 'UGLI2+3_nonEuropeans.txt' file. It is up to the researcher whether he/she wants to remove them or correct for population stratification in his/her genetic analysis.

9.2 Within-Europe populations

We performed a second analysis by creating the GRM, 20 PCAs, and PC-loadings of only the European 1000G population (503 samples) to assess the UGLI2+3 population structure within the European population. The same stratification and outlier identification method as described above were applied. Only the first two PCs were deemed informative. The Lifelines population is entirely based in the Northern-Netherlands, hence the expectation was that most Europeans will cluster with the central European population (CEU). We identified 61 Europeans that did not cluster with the CEU population (**Figure 10, right-side**). These were again not excluded from the analysis but are listed in the 'UGLI2+3_nonEuropeans.txt' file. It is up to the researcher whether he/she wants to remove them or correct for population stratification in his/her genetic analysis

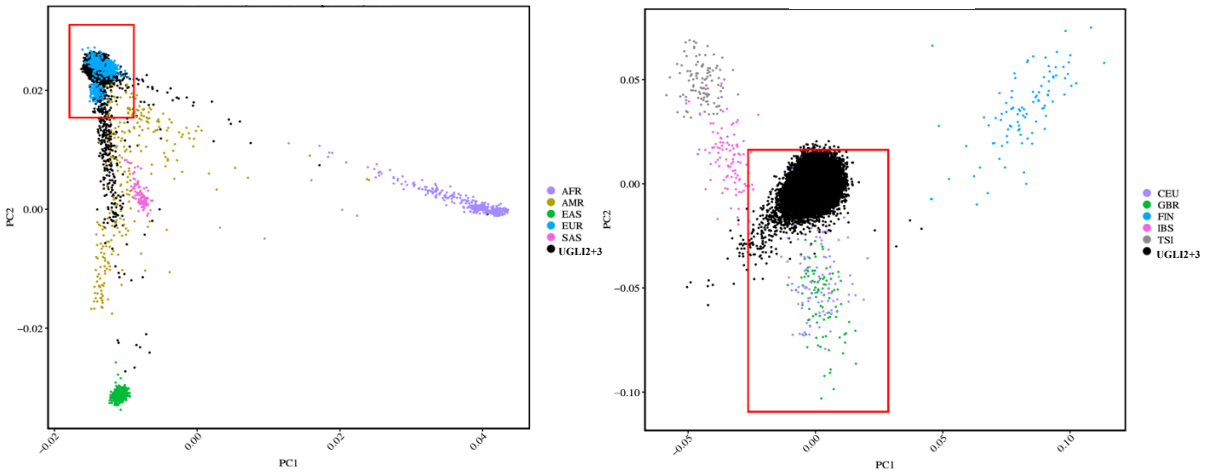

Figure 10: Population stratification analysis of the UGLI2+3 samples with the superpopulations on the left and within the European populations on the right. On the left, non-Europeans were defined as >4 SDs from the centroid (red square) of the 1000G European population (blue dots) using the first five PCs (300 non-European individuals). On the right, non-central-Europeans were defined as >4 SDs from the centroid (red square) of the European 1000G population (purple dots) using the first two PCs (61 non-CEU individuals).

10. Mendelian errors

After establishing the family relations within the combined set of CytoSNP (UGLI0), UGLI-GSA (UGLI1) and UGLI2+3 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 autosomal SNPs with more than 1% of Mendelian errors

across all Parent-Offspring (PO) pairs were observed and hence no autosomal SNPs were excluded at this step. For chromosome X we excluded 25 SNPs that exceeded the more than 1% Mendelian errors threshold.

11. Hardy-Weinberg equilibrium in unrelated individuals

Lastly, we re-calculated HWE p-values per SNP including only unrelated individuals within UGLI2+3. 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 5 (autosomal markers) and 7a (X chromosomal markers) in this QC protocol, but now on the subset of unrelated individuals for the autosomal markers and the unrelated females for the X chromosome markers, respectively. All genetic markers with a HWE p-value ≤ $1x10^{-6}$ were excluded (N=2,049 autosomal markers and N=458 X chromosomal markers) leaving 502,208 and 19,284 markers on the autosomal and X chromosome, respectively.

12. Alignment with HRC (imputation reference)

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, the positions of the genetic markers in the UGLI2+3 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+3 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) exclude SNPs with allele frequencies differing >10% from HRC; (v) update alleles to align with the positive strand; (vi) update position; and (vii) update reference and alternative alleles to match those on the HRC imputation server.

With this step 44,001 (8.80%) and 798 (3.81%) autosomal and chromosome X genetic markers, respectively, (41,176 unmapped, of which 34,009 indels; 1,320 palindromic; 1,207 non-matching alleles, and 299 with >10% allele frequency difference) were removed prior to imputation, leaving 456,566 autosomal markers and 20,127 X chromosomal markers in the final dataset.

13. Genetic imputation

A final set of 60,157 samples and 476,693 markers on autosomal and X chromosomes passing all QC steps described above and were used for genetic phasing and imputation. We previously used the official Sanger Imputation service for Genetic phasing and imputation that makes use of the Haplotype Reference Consortium panel (http://www.haplotype-reference-consortium.org). The HRC is a large whole-genome sequenced panel of almost 64,000 individuals with predominantly European ancestry combined from 20 different studies and contains more than 39 million SNPs. The online HRC imputation service makes use of the SHAPEIT2 or EAGLE2 phasing tools and the PBWT imputation tool.

We attempted to use this imputation service but noticed a low number of SNPs with a high imputation score (INFO score), possibly due to the PBWT imputation tool they use being relatively old (2014). Instead, we opted for an in-house phasing and imputation approach using a subset of ~11,000 HRC

samples as a reference panel with the EAGLE2 phasing and IMPUTE5 imputation tools. This resulted in the imputation of 39,131,940 variants. The distribution of the imputation qualities is shown in Figure No filters were applied on the final imputed dataset, because we believe it is up to the discretion of each individual researcher to apply their desired filters.

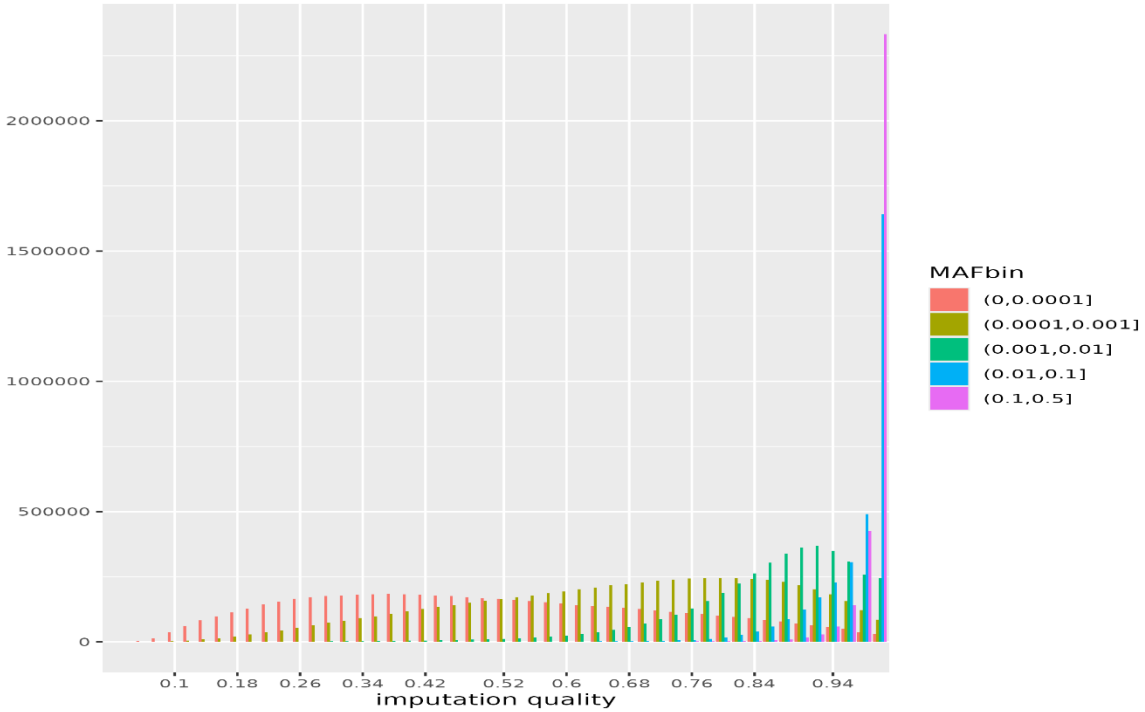

Figure 11: Distribution of the imputation qualities of the imputed (and genotyped) variants in UGLI2+3 per minor allele frequency bin (MAFbin).

Summarizing table

QC Steps		Variants								Samples			
		Autosomes + XY				ChrX							
		Remaining	Excluded	%	Flagged	Remaining	Excluded	%	Flagged	Remaining	Excluded	%	Flagged
Pre-QC (including chr Y and MT)		615682	0	0%	-	22346	0	0%	-	63553	0	0%	-
Variant calling	DQC = 0.82	615682	0	0%	-	22346	0	0%	-	63141	412	0.65%	-
	QC call rate = 0.90	615682	0	0%	-	22346	0	0%	-	62826	315	0.50%	-
	Plate exclusion cut-off = none	615682	0	0%	-	22346	0	0%	-	62826	0	0%	-
Merging variants and allele frequency concordance > 0.95		615575	107	0.02%	965	22346	0	0%	29	62826	0	0%	-
Exclude chr Y and MT		614532	1043	0.17%	-	22346	0	0%	-	62826	0	0%	-
Duplicate marker and sample filtering		613549	983	0.16%	-	22337	9	0.04%	-	62540	286	0.46%	-
Sample & variant Call rate filtering	Low = 0.80	613483	66	0.01%	-	22336	1	0.004%	-	62540	0	0%	-
	High = 0.95 (sam) + 0.98 (var)	600654	12829	2.09%	-	22055	281	1.27%	-	62111	429	0.69%	-
MAF < 0.003% (= three or less minor alleles) filtering		544959	55695	9.27%	-	20433	1622	7.94%	-	62111	0	0%	-
HWE p < 1e-10 filtering		528709	16249	2.98%	-	20248	185	0.91%	-	62111	0	0%	-
Sample heterozygosity filtering	R1 = remove 7 plates	528709	0	0%	-	20248	0	0%	-	61704	408	0.66%	-
	R2 = remove other outliers	528709	0	0%	-	20248	0	0%	-	60846	858	1.39%	-
Sample mix up identification (pedigree check)		528709	0	0%	-	20248	0	0%	-	60377	469	0.77%	-
Sex check (mismatch)		528709	0	0%	-	20248	0	0%	-	60157	220	0.36%	-
Array, batch, and plate effects check	Array effects	504694	24016	4.54%	-	19816	432	2.13%	-	60157	0	0%	-
	Batch + Plate effects	504257	2563	0.51%	38560	19767	49	0.25%	1119	60157	0	0%	-
Population Stratification (Global)		504257	0	0%	-	19767	0	0%	-	60157	0	0%	300
Mendelian Errors Check 100 errors		504257	0	0%	-	19742	25	0.13%	-	60157	0	0%	-
HWE p < 1e-6 in unrelateds		502208	2049	0.41%	-	19284	458	2.30%	-	60157	0	0%	-
Move chrXY from autosomes to chrX		500567	1641	→	-	20925	-	-	-	60157	0	0%	-
Liftover to GRCh37 and Alignment to HRC		456566	44001	8.80%	-	20127	798	3.81%	-	60157	0	0%	-