UGLI - GSA (release 2.0) Quality Control Report

Updates in release 2.0:

- Genetic variants have been compared with the global reference panel of the Haplotype Reference Consortium and consequently an additional 23,391 variants were removed prior to imputation (see section 10, page 20).

The **U**niversity Medical Center of Groningen **G**enetics **L**ifelines **I**nitiative **(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 38,030 participants assessed using the Infinium Global Screening Array® (GSA) MultiEthnic Disease Version 1.0. In this QC screening we included all genotyped samples, but we focused on QC of genetic markers on the autosomes and chromosomes X (N=691,072 markers).

In brief, first we made translations and corrections specific from the GSA platform to a general context of usage; namely, strand harmonization and removal of duplicate markers within the array. Secondly, low quality samples and markers were carefully filtered with a two-steps procedure of call rate thresholding. Further possible genotyping errors were assessed at the marker level by detecting variants that deviated significantly from Hardy-Weinberg equilibrium (HW) and 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 and Y chromosomes, 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). Moreso, to further evaluate sample mix-ups we compared the concordance of genotype calling among a subset of samples with genotype information from a different array (CytoSNP 250k array, n=606, from the Lifelines GWAS data set) and whole genome sequence (WGS, n=143, from the Genome of the Netherlands (GoNL) project). Subsequently, we ascertained Mendelian errors and further removed variants that deviated from Hardy-Weinberg equilibrium (HWE) in unrelated individuals. Finally, population stratification was inspected by a principle components analysis (PCA), incorporating samples from 1000 Genomes (1000G) and GoNL projects. 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. Furthermore, the code and detailed description of the process can be found in: https://github.com/molgenis/GAP.

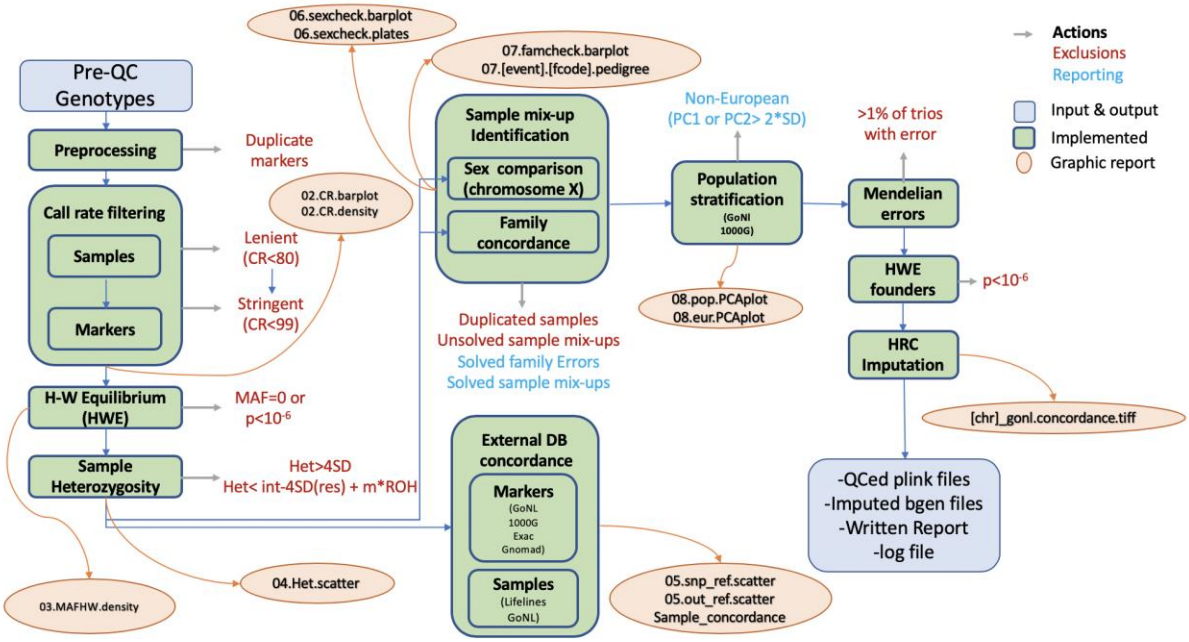

Figure 1. Steps and metrics evaluated in the quality control of the UGLI genotype data.

Step-wise quality control

1. Pre-quality control steps

1a. Illumina intensity files (IDAT) were used to harmonize strand direction according to Illumina manifest, and to determine genotype calls using Opticall (https://opticall.bitbucket.io), which algorithm was designed for genotype calling of genetic markers, in particular for rare alleles [minor allele frequency (MAF)<1%)](1).

1b. Harmonized genotypes were converted to PLINK (https://www.cog-genomics.org/plink/1.9/) files (BED) and separated into chromosomes (autosomal 1-22, X, Y, pseudoautosomal XY and mitochondrial MT) to be further processed.

2. Filtering by call rate and built-in duplicate markers

For autosomal and pseudo-autosomal chromosome we removed built-in duplicate markers and markers with high missing rate using a two-thresholds two-steps process: first by samples and then by variants, filtering first with a lenient missing rate threshold (20%) and then by applying a more stringent missing rate threshold (1%). All the steps

here were done with PLINK v1.9b3.32 commands `--missing` `--remove` and `--exclude`, following this workflow:

2a. Calculate missing rate per individuals.
2b. Identify duplicate markers by positions and allelic content (in such a way that tri-allelic markers would make two different markers). For duplicate markers, the one with the highest missing rate is flagged for removal.
2c. Remove duplicate markers identified in step 2b, and also remove samples with missing rate higher than 20% determined in step 2a.
2d. Calculate missing rate for markers
2e. Remove markers with missing rate higher than 20%
2f. Recalculate missing rate for individuals
2g. Remove samples with missing rate higher than a certain stringent missing rate (1%)
2h. Recalculate the missing rate for markers
2i. Remove markers with missing rate higher than certain stringent missing rate (1%)

We identified and removed 556 duplicated (by position and allele) markers. After the lenient call rate filter (80%, i.e. missing rate=20%), the distribution of call rates is normal-like around the 99.5%-99.8% marks (Figure 2). Therefore, we decided for a stringent call rate threshold of 99%. After call rate filtering 36,930 (97.1%) samples and 641,303 (95.3%) markers remained (Figure 3).

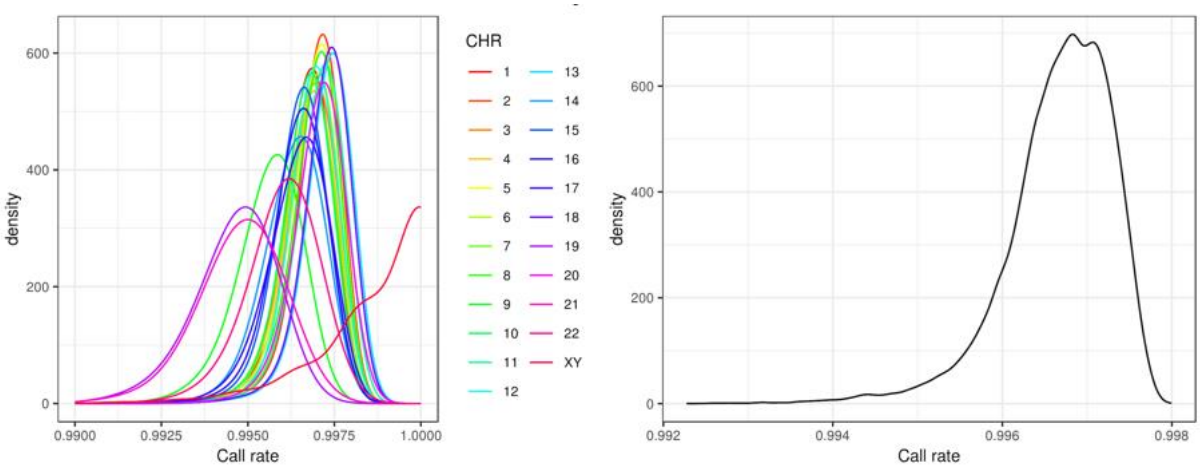

Figure 2. Call rate distribution for markers after lenient (>80%) filtering. Left: call rate distribution by chromosome. Right: call rate distribution of all the remaining markers.

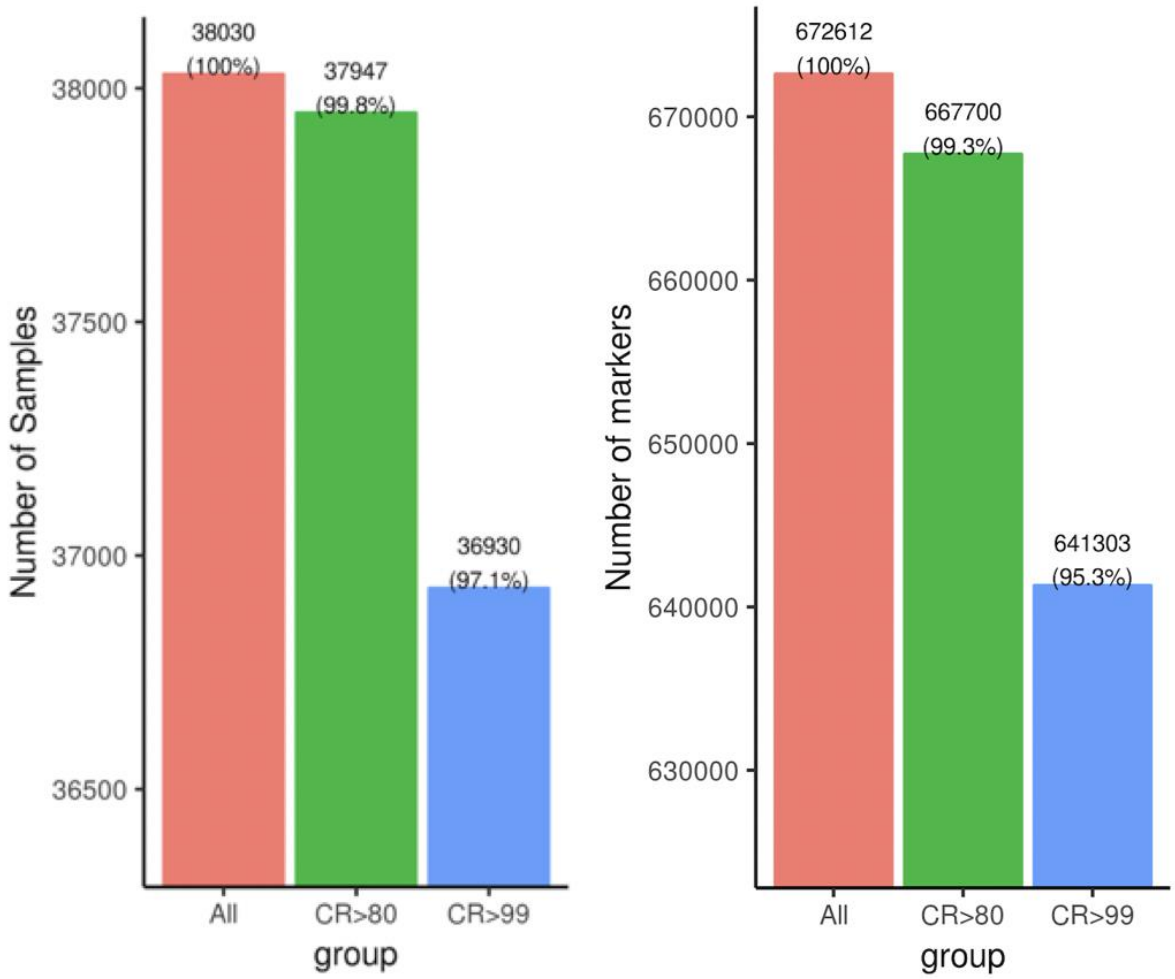

Figure 3. Call rate filtering for GSA genotyping of 38,030 samples after removal of built-in duplicated markers. Left: samples. Right: autosomal and pseudoautosomal markers

3. Minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HW)

We calculated the allele frequencies (Figure 4) and the HW (Figure 5) using PLINK v1.9b3.32 commands `--freq` and `--hardy`. Monomorphic markers (MAF = 0) and/or markers with a HW p-value $\leq 1x10^{-6}$ were considered uninformative and of poor quality, and were removed. For the HW test no pedigree information was available yet, so a lenient threshold is used. This HW QC step is repeated after establishing family relations of all samples (see step 8).

A total of 75,033 markers were found to be monomorphic and another 7,874 were outliers for HW. These were removed in this step.

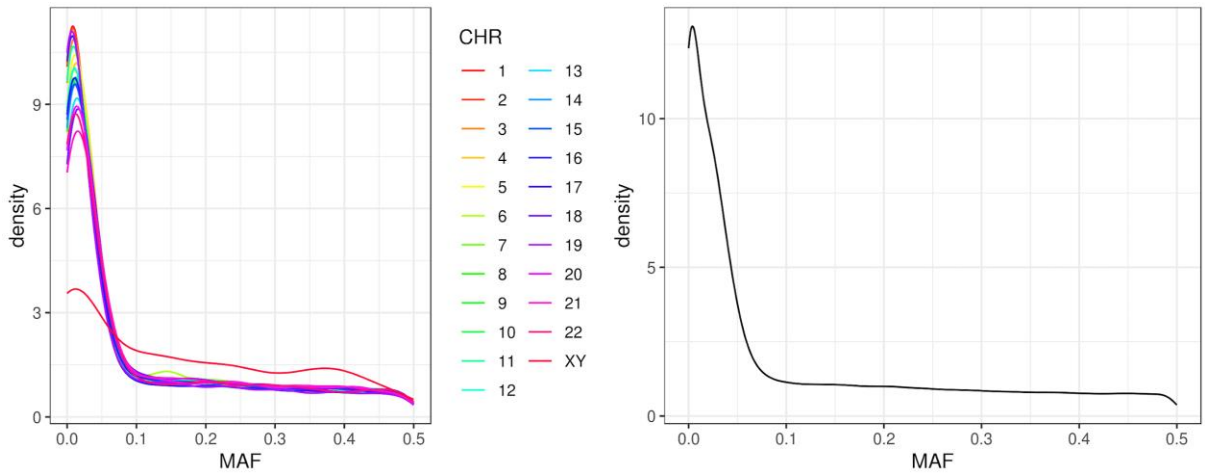

Figure 4. MAF distribution for markers. Left: by chromosome. Right: all remaining markers.

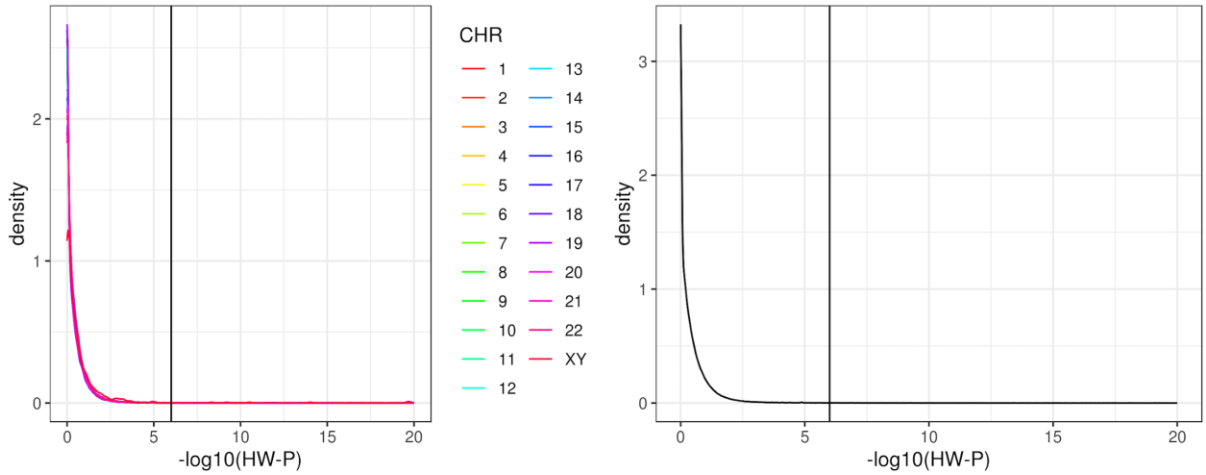

Figure 5. HW log(p-value) distribution for markers. Left: by chromosome. Right: remaining markers together.

4. Samples 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 contamination or issues during genotyping process. To calculate heterozygosity we filtered out the HLA region (to avoid inflating the heterozygosity measured by LD) in the chromosome 6 and merged all chromosomes aftering selecting independent variants (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 excluded 194 samples that were considered as heterozygosity outliers, Figure 6 shows the exclusion thresholds as red lines. To further understand if these heterozygosity outliers were being driven by a higher missing call rate, we tested if heterozygosity was associated with missing rate levels. However, there seems to be no relationship (**Figure 7**).

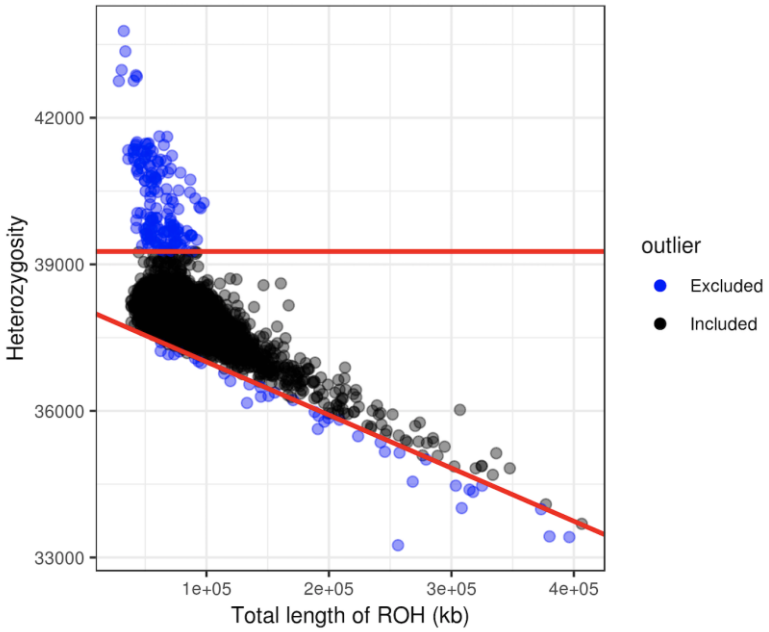

Figure 6. Heterozygosity and ROH of remaining samples. Red lines represent the filtering thresholds. Blue dots represent samples that are excluded based on more than 4 standard deviations (SD) above the mean heterozygosity or more than 4 SD of the residuals of the linear regression between heterozygosity and ROH below the predicted heterozygosity from this same linear regression analysis.

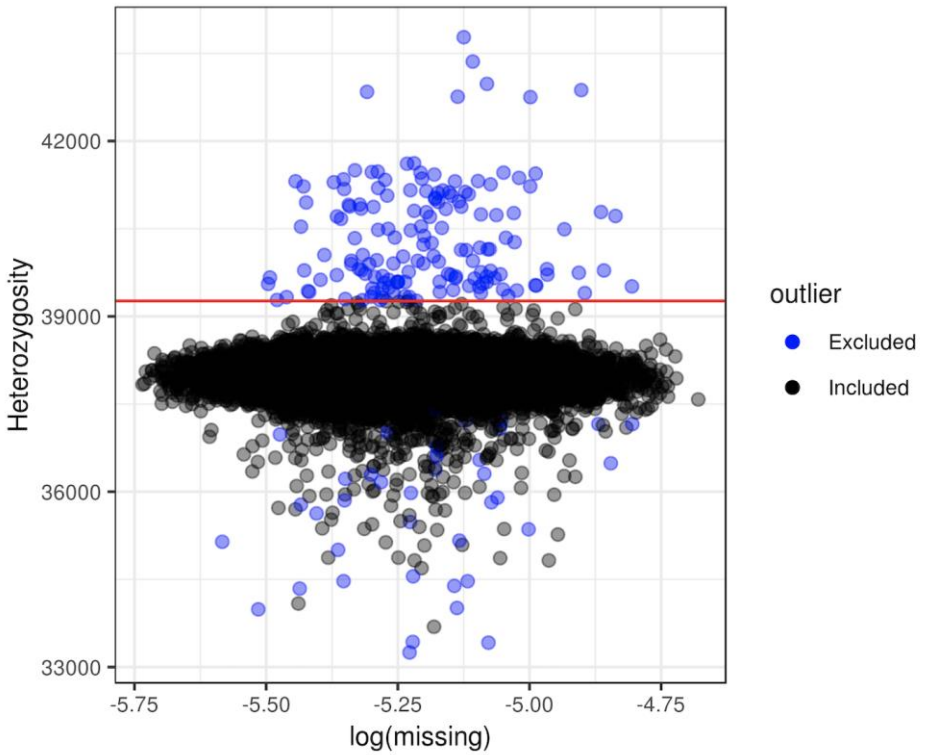

Figure 7. Relation between heterozygosity and missing rate. Blue dots are considered as heterozygosity outliers.

5. Sample mix-ups

Sample mix-up is investigated by looking at gender mismatch, where gender information of each sample 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. We took advantage of the pedigree information and familial relationships between Lifelines samples and of the ability to use statistical models to ascertain sample mix-ups by comparing genetic relations of each sample with the expected genetic sharing from pedigree information. We did these analysis in parallel and each potential sample mix-up detected was carefully analysed and evaluated taking into consideration plate number and position as well as the supposed volunteer's questionnaire information regarding first-degree relationship (children, partner, parents, and siblings) with other Lifelines members. The specific details on the gender mismatch and familial relationship concordance analysis are described below.

5a. Chromosome X QC and sex check

The chromosome X was analysed independently from the other chromosomes. We analyzed all samples that passed QC at this level of

filtering (step 4). At the marker level, we first applied the same thresholds as for the autosomal chromosomes in step 2 (i.e., filtering by call rate (N=1327, 7.4%) and built-in duplicate markers (N=12)). Next we inferred genetically determined sample sex by calculating heterozygosity of chromosome X with PLINKv1.9b3.32 (`--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 analysis.
After full sex and familial information was ascertained we filtered chromosome X to remove monomorphic markers (N=2,780) and HW outliers ($p<1x10^{-6}$) with only parental females (N=35 markers).

5b. Pedigree analysis

We used individual genetic information to infer the relationship between each possible pair of participants. We compared this with the pedigree information available from the Lifelines database, which was optimized during sample selection. The genetic relationship between participants was inferred using KING 2.2 (http://people.virginia.edu/~wc9c/KING/) with the commands `--relations --degree 2`. 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 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.

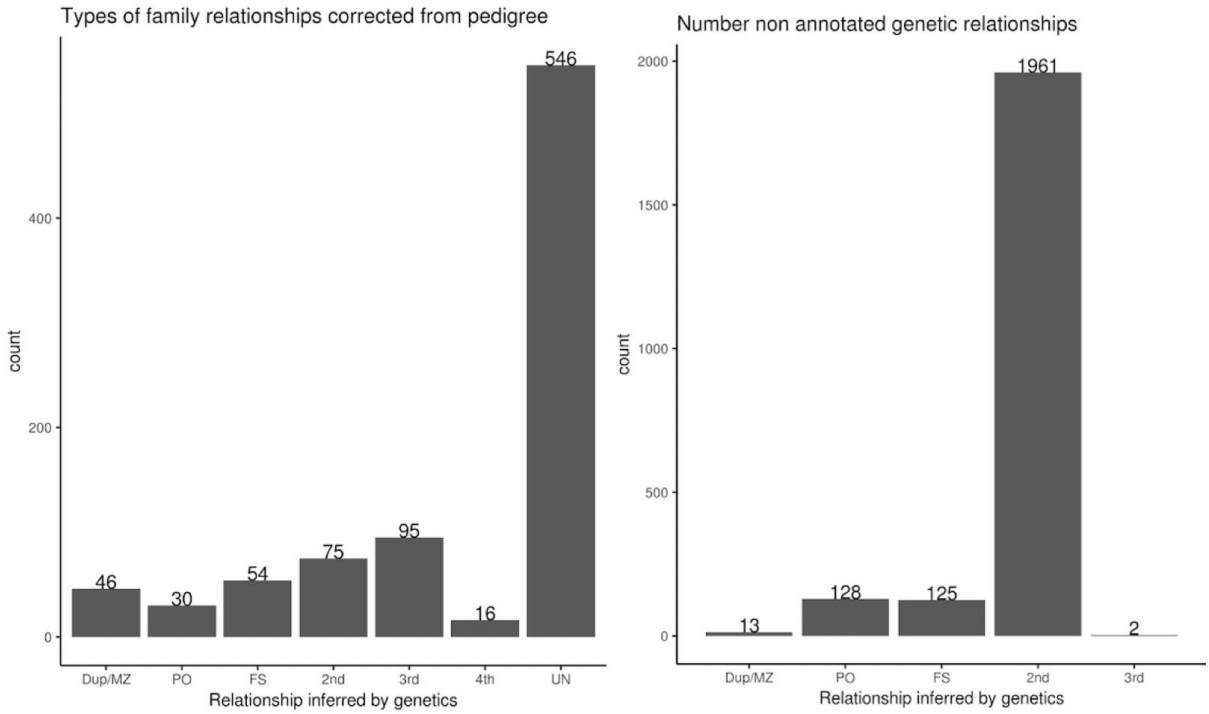

Figure 8. Summary of the genetic relationships calculation. Dup/MZ: duplicates /monozygotic twins, PO: parent-offspring, FS: full siblings, UN: unrelated, and ordinal numbers indicate relationship degree. Left: Family relationships flagged as errors, calculated genetic relationships are shown. Right: relationships not indicated by the family information and found with genetic calculation.

A total of 71,134 known family relationships were confirmed, while 862 (1.2%) relationships were flagged as errors, and an additional 2,229 new relationships were found. We analysed any family relationship flagged as "error" that resulted in a genetically calculated first degree or "unrelated" relationship (**Figure 8**, left), as well as all first degree relationships (Monozygotic twin / duplicates, Parent-offspring or Full siblings) that was not reported in the Lifelines pedigree (**Figure 8**, right). For each of the families involved in one or more of these events we visualized the information in a pedigree plot, and we coupled it with the age, sex and questionnaire information (see example in **Figure 9**). 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.

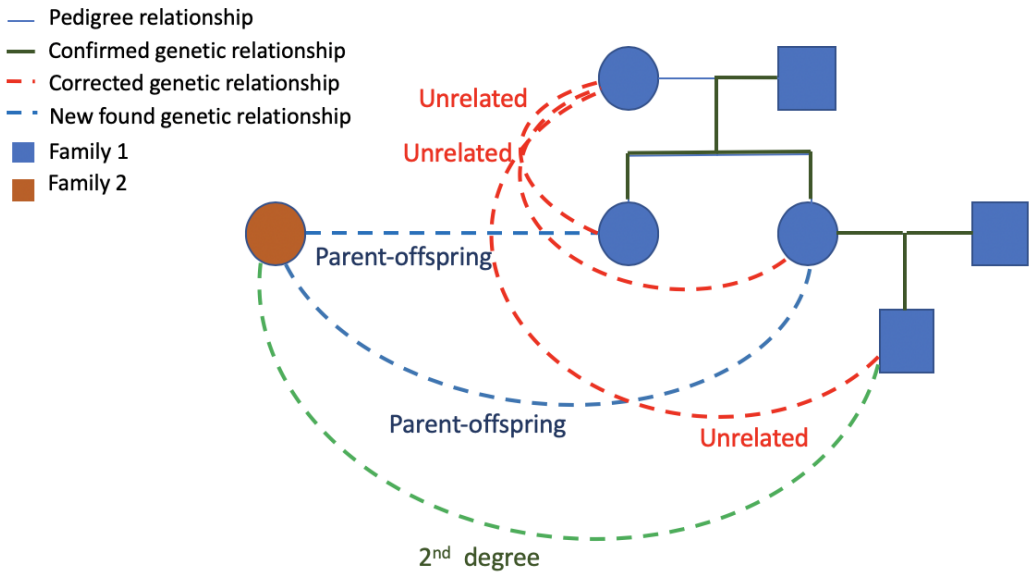

Figure 9. Example pedigree analysis of a sample mixup. The participant from family 2 (in gold) is actually the grandmother of family 1, while the supposed grandmother of family 1 (in blue) does not belong to this family.

A total of 473 samples had one or more of the events shown in **Figure 8**. Of these samples, 60 (12.7%) were in the event of real monozygous twins, 32 (6.7%) were distant relationships marked in the Lifelines pedigree as "unrelated", 48 (10.1%) were corrections of familial information within a family, 97 (20.5%) were new relationships found, and 236 (49.9%) were identified as sample mix-ups. Of these sample mix-ups, 56 (23.7%) 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 (180) were excluded. Steps 2 to 5b of the quality control process were repeated in the new set of QCed samples (during which an additional 4 samples were identified to be of poor quality due the second filter of call rate and 16 additional samples were identified as outliers for heterozygosity and therefore removed in this process) to verify that no additional sample mix-ups were present after this correction process. As shown in **Figure 10**, no familial errors between first degree relatives or new findings were observed, although given the merging of big families, new distant relationship were incorrectly identified as unrelated relationships in the Lifelines pedigree.

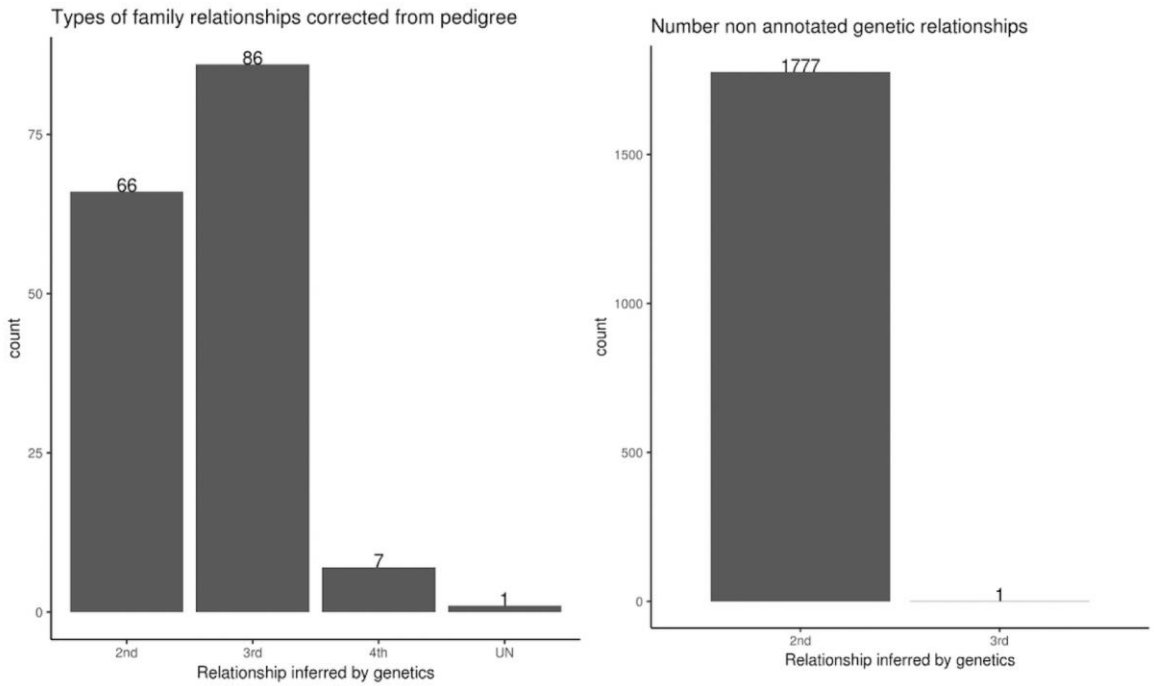

Figure 10. Summary of genetic relationships calculation in the corrected population. Dup/MZ: duplicates /monozygotic twins, PO: parent-offspring, FS: full siblings, UN: unrelated, and ordinal numbers indicate relationship degree. Left: Family relationships flagged as errors, calculated genetic relationships are shown. Right: relationships not indicated by the family information and found with genetic calculation.

The samples still flagged as “Non-concordant” by sex and classified as mix-ups from the pedigree were visualized in the plates to try to identify patterns and possible sources for the mix-up. Samples in plates with a high number of mix-ups (>10 mix-ups) that were near the wells of the mix-ups (nearby samples) were considered suspicious, as sample mix-up errors were more likely to have occurred there. We used familial information to evaluate their genuinity. If there were no relationships to check the identity of these nearby samples, they were also considered possible mix-ups and removed.

After this step we removed 42 samples still flagged as “Non-concordant” by sex, 5 samples flagged as “Failed sex imputation” and with no relationship from relationship calculation in the Lifelines pedigree file, and 2 additional nearby samples without relationships in the relationship calculation in the LifeLines pedigree[1].

[1] If a sample belonged to a family according to the Lifelines pedigree file and the only relationship was "unrelated" (UN) in the KING file, the sample was still not removed. In this sense, UN is still considered a pedigree relationship. These mostly concern spouses that are genetically unrelated to the other family members. This means some samples (between 2-30), that failed sex imputation and might have UN as the only relationship, were not removed at this point. This will be reviewed in a later release.

6. Population Stratification

In this step we aimed to identify non-European individuals. Using genotype data from UGLI, 1000 Genomes (1000G), and GoNL parental individuals we built a joint data set, with the markers present in all three data sets. The variants in this combined dataset were filtered by MAF>10% and call rate of >99%; additionally the HLA region of chromosome 6 was removed. This dataset was further pruned using PLINK with the `--indep` option and the parameters: windows: 1000, step: 5, r^2 threshold: 0.2. This resulted in a set of 43,587 markers that were used for further analysis. We then calculated the principal components using only the participants from 1000G and GoNL dataset and projecting UGLI participants into them (**Figure 11,** left), using the PLINK commands `--within -pca --pca-cluster-names`. Next, using the first two principal components we flagged as Europeans all individuals clustering with GoNL or 1000G European populations or no more than 3 SD away from their extremes according to both PC1 and PC2 (**Figure 11**, right). A second PCA run was carried out using the same steps but including only the UGLI participants flagged as European, GoNL and the European populations of 1000G, and more stringent filters (2 SD) was used to flag Europeans this time.

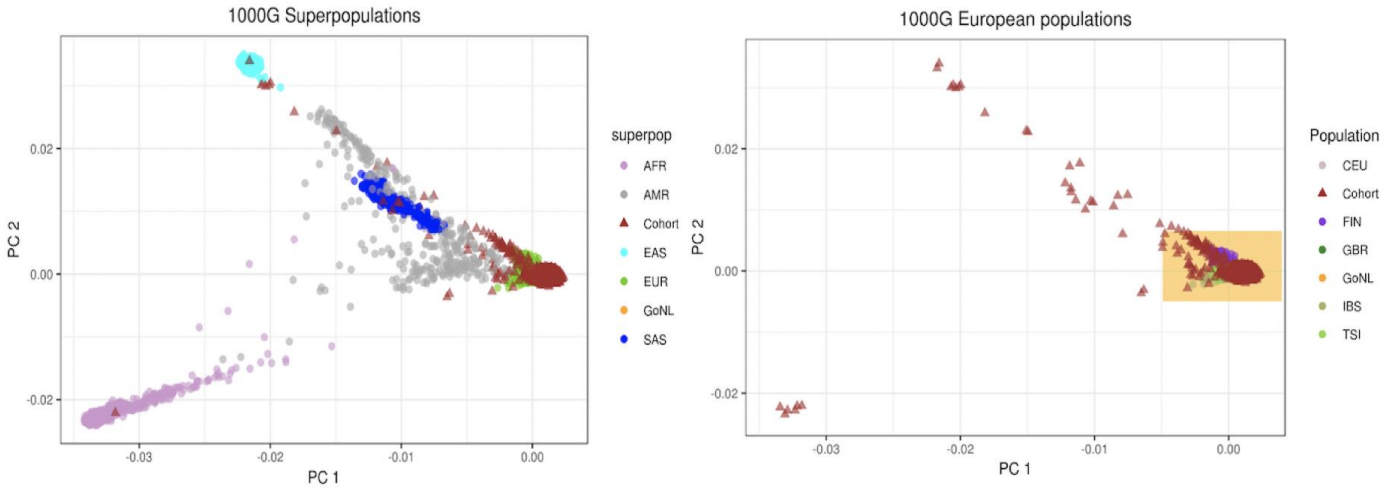

Figure 11. First PCA analysis: participants of the UGLI cohort projected in the principal components of 1000G (all populations) and GoNL. Left: UGLI cohort, GoNL, and 1000G superpopulations. Right: Same analysis, but showing only UGLI cohort, GoNL, and 1000G European populations. The light orange square indicates the 3 SD threshold used to flag the European individuals. AFR: African samples from 1000G; AMR: American samples from 1000G; EAS: East-Asian samples from 1000G; EUP: European samples from 1000G; SAS: South-East Asian samples from 1000G. CEU: Caucasian European samples from Utah from 1000G; FIN: Finnish samples from 1000G; GBR: Great-British samples from 1000G; IBS: Iberian samples from 1000G; TSI: Toscan samples from 1000G.

From the first PCA analysis we detected 35 samples that did not qualify as Europeans (**Figure 11**, right), the second PCA analysis with only European samples showed no more UGLI participants with a non-European ancestry (**Figure 12**). We can also note that UGLI population clusters partly with GoNL and British in England and Scotland (GBR). Given the limited number of non-European samples, we did not remove them,

but kept these in the final file and used them for imputation. We however provided a file with indication of European ancestry as this could be useful to researchers depending on their research question.

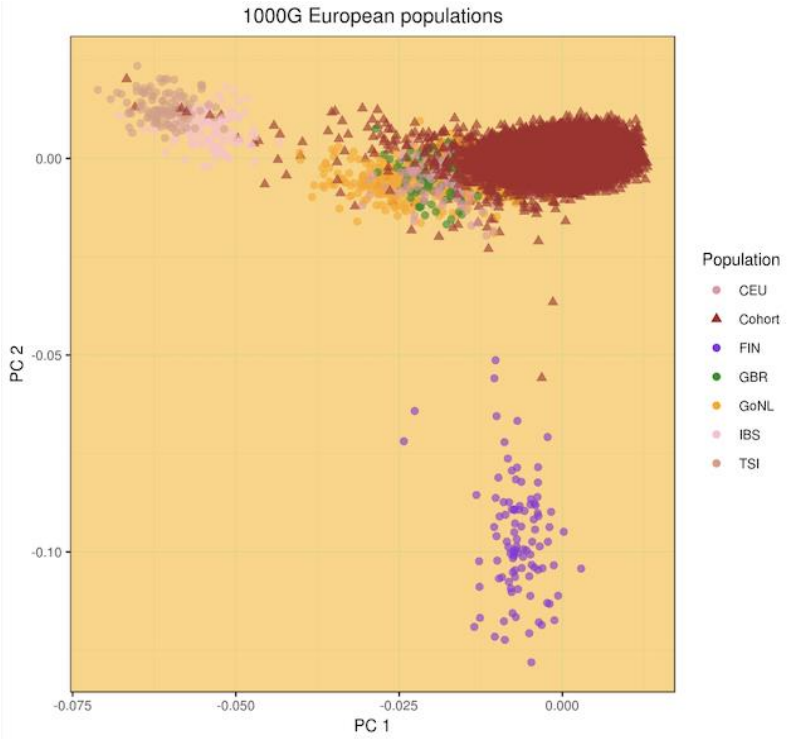

Figure 12. Second PCA analysis: European participants of the UGLI cohort projected in the principal components of 1000G (only European populations). The light orange square (ranging in the whole scale) indicates the 2 SD threshold used to flag the European individuals.

7. External concordance

We used external information when possible to evaluate and observe possible deviations from already published QC reports from other European cohorts. The external concordance can be grouped in two different strategies: i) MAF concordance, which evaluates the observed MAF across markers in UGLI and the MAF reported in other European cohorts. And ii) intra concordance, which evaluates the concordance of genotype calls across a set of UGLI participants that have been genotyped using other arrays or WGS. To evaluate the concordance we used our own workflow in R v3.5.1 and the package SNPstats.

7a. MAF concordance with external cohorts

We evaluated the concordance with the MAF of markers passing all previous QC steps with MAF previously reported by external large scale genomic European cohorts: GoNL, 1000G, Exac and gnomAD. We defined as a measurement of concordance the relative percentage of variants that deviated at most 4 SDs of the residuals calculated from a linear model between MAF reported by UGLI (y-axis, Figure 13, left and right) and the MAF reported by an external cohort (x-axis, Figure 13, left and right). Across all external cohorts the

maximum percentage of discordance was 0.76% with the ExaC dataset, however this percentage dropped to 0.68% when we include only non-Finnish Europeans (**Figure 13**). The lowest percentage of discordance was found with the European populations of 1000G at 0.37%. These represent similar percentages as the ones reported by UKBiobank (3) (~0.3%)

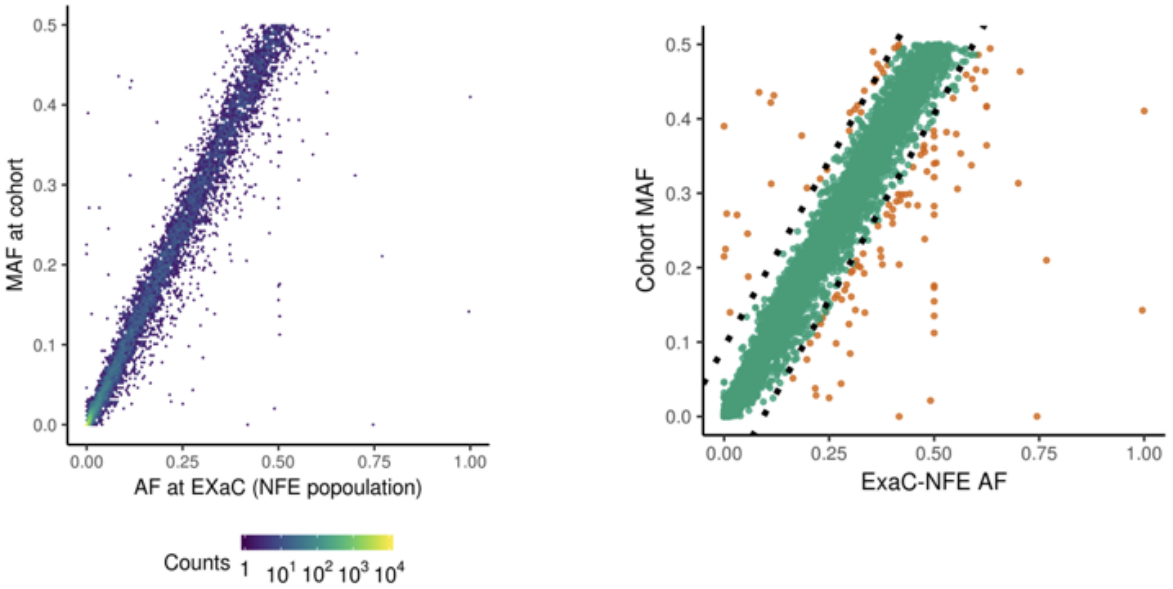

Figure 13. MAF concordance with external cohorts. On the left, hexagon plot showing the concordance between the MAF reported in UGLI (y-axis) and the one from ExaC (non-Finnish Europeans) on the x-axis. Color represents density of markers per hexagon. On the right, a scatter plot showing the same as left, but orange dots outside of dotted lines are identified as discordant markers.

7b. Intra concordance

Some of the UGLI participants have also been part of other genotyping initiatives such as Lifelines DEEP and GoNL. Therefore we also evaluated the concordance of genotype calls using the GSA array, against the cytoSNP array used in the Lifelines GWAS project and WGS performed in the GoNL project. This intra-concordance was ascertained for each sample that was present in two data sets as the percentage of markers that had the same genotype in both datasets. We observed that for the 606 participants genotyped by both GSA and the CytoSNP array (**Figure 14**, left), the concordance was very high (mean concordance ≥ 99%, minimal concordance = 99.82%) as well as for the 91 participants who were also genotyped in GoNL (mean concordance ≥ 98%; minimal concordance = 98.15%,) (**Figure 14**, right). None of the samples showed a concordance below 95%, therefore all were retained in the dataset.

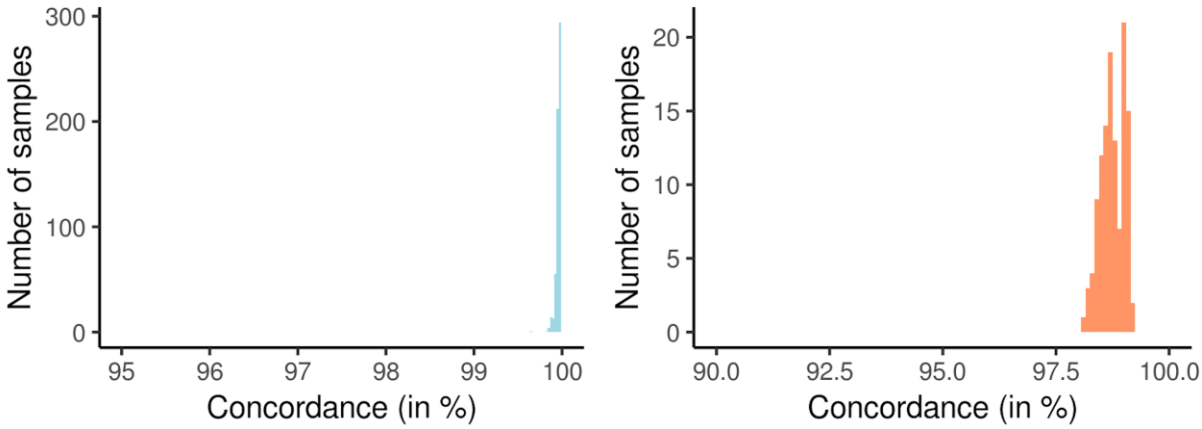

Figure 14. Intra-concordance at sample level: On the left, distribution of percentage of sample level concordance in the 606 samples genotyped by the GSA array and the Cytochip. On the right, distribution of percentage of sample level concordance in the 91 samples genotyped by the GSA array and by WGS.

We also evaluated the intra-concordance at a SNP level. This meant that for each SNP that was overlapping we calculated the relative percentage of concordant genotypes across all the samples that were shared across the studies. When comparing with the genotypes called using the Cytochip we observed that across the 51,362 variants that these two platforms shared the mean concordance ≥ 99% (**Figure 15**, left). For the 236,081 variants that were shared between the GSA array and the GoNL WGS, a mean concordance ≥ 95% was observed (**Figure 15**, left). Because we have more trust in the genotype data from the GSA chip, no variants were removed at this stage.

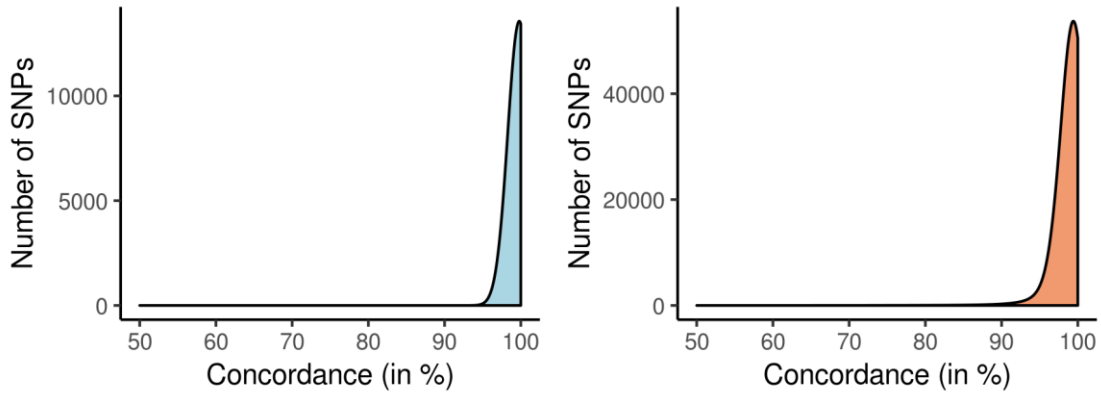

Figure 15. Intra-concordance at SNP level. left: UGLI vs LLDEEP, Right: UGLI vs GoNL.

Similarly to the previous section MAF concordance with external cohorts, we also compared the MAF reported by the shared variants between UGLI and the Cytochip and WGS. As expected from the SNP level concordance, we observed that nearly all variants between both arrays (GSA and GWAS) had almost the same MAF (**Figure 16**, top plots in light blue). Yet, for the variants overlapping

with WGS, we found that concordance decreases as the variants became more rare (lower MAF) (**Figure 16**, bottom plots in orange). This was expected since in GoNL the sequence depth was not very high, therefore plausible errors when calling rare genotypes are expected.

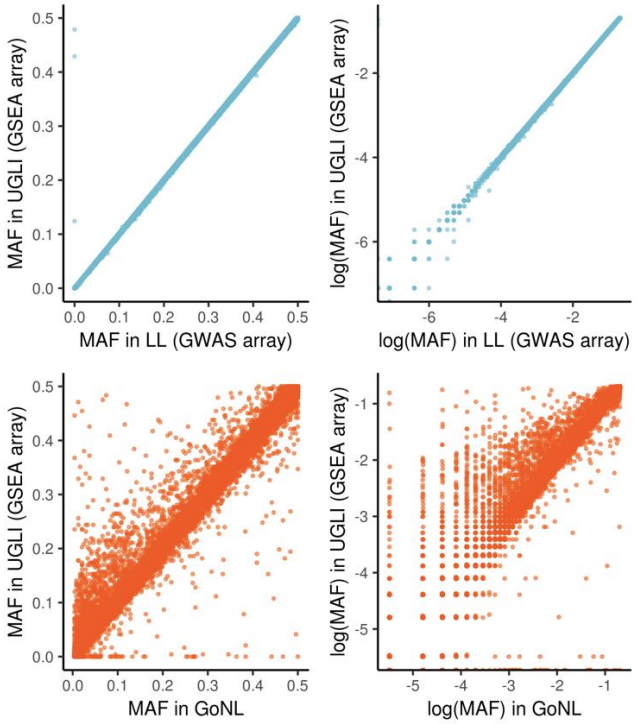

Figure 16 MAF intra-concordance: On the top, scatter plots showing the concordance between the MAF reported in UGLI with the GSA array in the y-axis and the one reported using the Cytochip on the x-axis, with the upper right plot showing the log transformation of the MAF to better appreciate variants with low MAF. On the bottom, scatter plots showing the concordance between the MAF reported in UGLI with the GSA array in the y-axis and the one reported using the WGS on the x-axis, with the lower right plot showing the log transformation of the MAF to better appreciate variants with low MAF.

8. Mendelian errors and HW in founders

As a final step, we quantified the number of mendelian errors detected per each of the variants. 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 times we observed an error for each SNP. We excluded variants with more than 1% of Mendelian errors across all Parent-Offspring (PO) pairs (**Figure 17**, left). Using this threshold 2 variants were removed from UGLI.

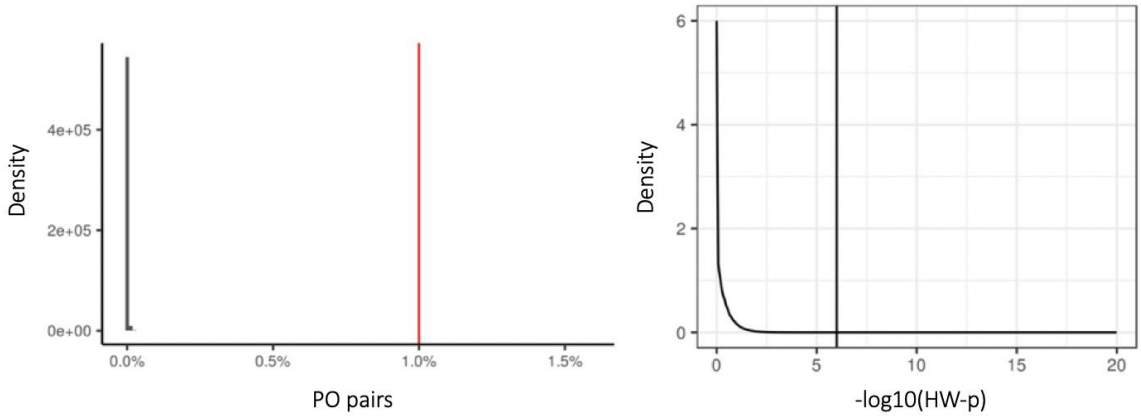

Figure 17. Distribution of variants. Left: percentage of PO pairs with errors, red line indicates the threshold to consider outliers (1%). Right: HW p-value, vertical line indicates the threshold.

Lastly, we re-calculated HW per SNP including only the founders of each pedigree within UGLI. Same as in step 2 in this QC protocol, we used PLINK and the command `--hardy`, however this time we include as input in PLINK the pedigree information (the .fam file) to consider only the founders. We excluded all variants with a HW p-value ≤ $1x10^{-6}$ (N=727) (**Figure 17**, right).

9. Batch differences

The UGLI samples were genotyped in 31 different batches of ~1200 samples during the period of ~5 months in two different laboratories (16 batches were processed in the Genetic Laboratory Erasmus MC Rotterdam, and 15 in the laboratory of the department of Genetics at UMCG (Groningen)). During this period each batch was analysed separately as the data became available. This allowed us to i) have a measure per batch of each of the parameters of the quality control steps, ii) identify problems of the genotyping process before it was fully finished, and iii) address these issues before the genotyping was finished.

While processing the batches, we found and removed 148 samples that were duplicated on purpose during the genotyping process as a quality control. We also found an unusual number of duplicates (or monozygous twins) in batch 4 (**Figure 18**). Furthermore, we noted that all these duplicate samples from batch 4 were part of plate 246, with a duplicated sample in the same well positions in plate 252. We hypothesized that one of the plates was a duplicate of the other. To confirm this and identify which of the plates was the duplicated one, we compared the sex information from self-reported information of each plate and compared it with the sex derived from genotype (**Figure 19**).

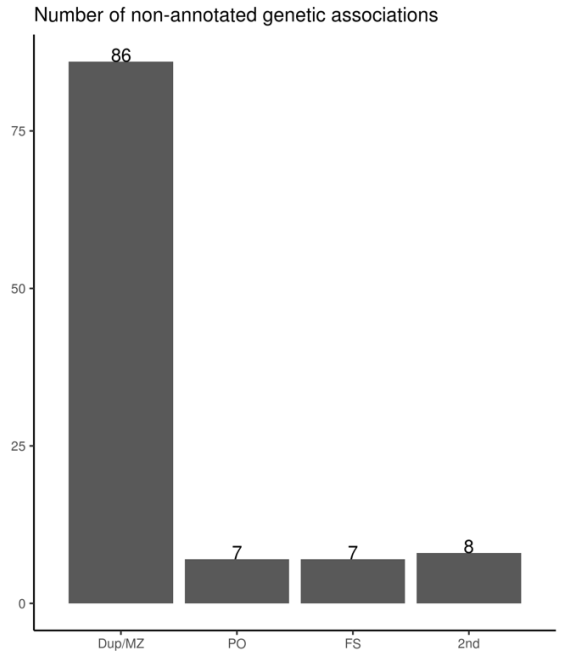

Figure 18. KING output of new found relationships for batch 4.

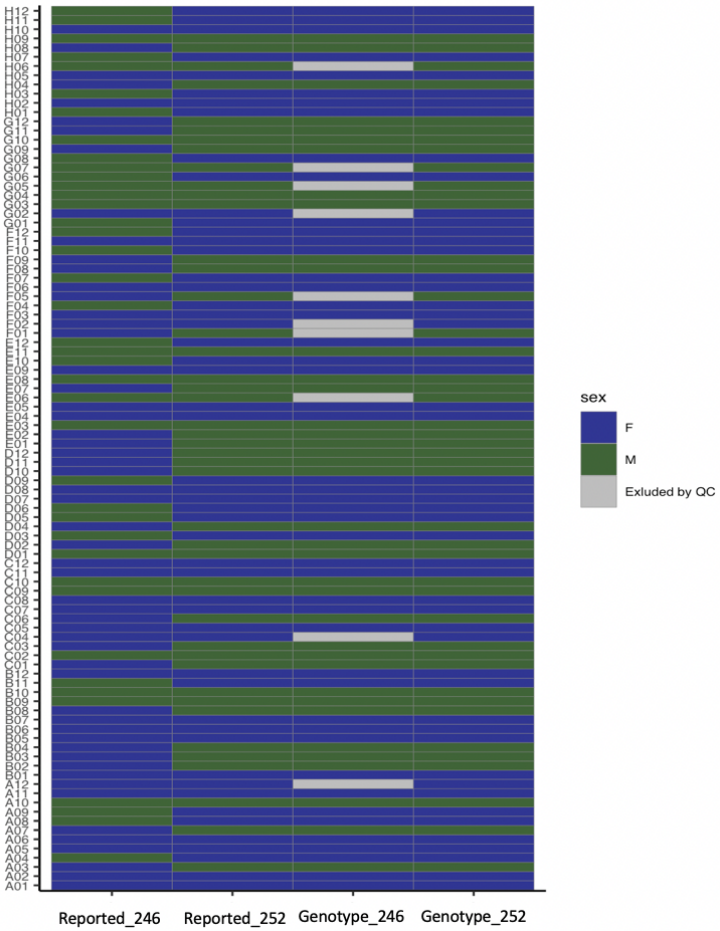

Figure 19. Sex information for each sample of the duplicated plates (246 and 252). Left: self reported sex information, Right: sex information calculated from genotype.

We found that the pattern for genotype derived sex information in plate 246 was reflecting the pattern of self-reported sex of plate 252 instead of the pattern of self-reported sex of plate 246. Therefore we determined that plate 246 in batch 4 was a duplicate of plate 252 in the same batch. This was reported back to the laboratory and then plate 246 was genotyped again with the correct set of samples. No other batches showed important deviations for the QC parameters.

When all the batches were available and had Opticall genotype calling information (after step 1), we merged all samples and then processed this complete dataset for quality control (steps 2-8). We compared the QC results visually between the individual batches and the merged population. Overall we found no significative differences from this comparison.
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 296 and 352), but these didn't seem to be attributable to the batch (**Figure 20)**. Finally **Figure 21** shows the PCA clustering of each individual batch. We observed that all the batches cluster together with a similar shape.

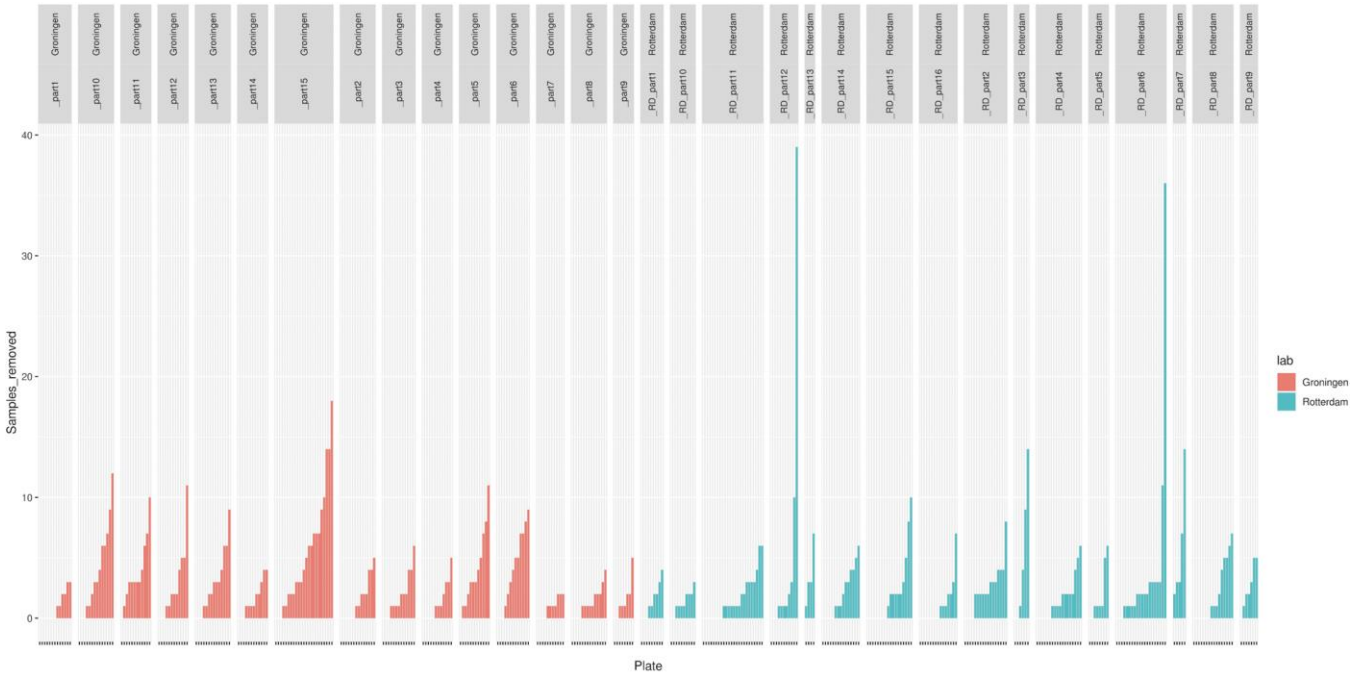

Figure 20. Samples removed for having <99% call rate by plate, batch and laboratory.

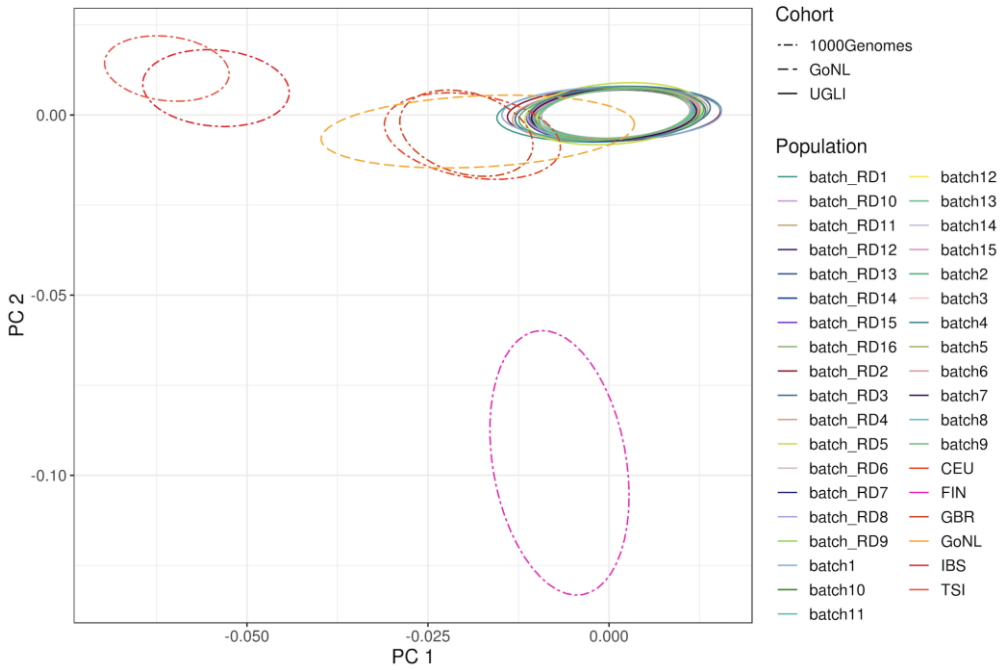

Figure 21. PCA clustering for each batch, GoNL and 1000G European populations.

10. Comparison with the Haplotype Reference Consortium reference panel

As a last step prior to imputation the genetic markers were checked against the global panel of the Haplotype Reference Consortium using the HRC-1000G-check-bim tool from McCarthy's group (McCarthy Tools (ox.ac.uk)). This tool checks whether all genetic variants

- are present in the HRC reference panel,
- are single nucleotide polymorphisms (and not insertion/deletion polymorphisms)
- are located on the same position as in the reference panel,
- have the same alleles as in the reference panel,
- similar frequencies (<10% difference) as in the reference panel,
- are not palindromic with an allele frequency >40%.

For variants passing these criteria (n=548,029), it codes them in such a way that they have been aligned to the positive DNA strand and have the same reference and alternative alleles as in the reference panel.

11. Genetic Imputation

A final set of 36,339 samples and 548,029 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. The dataset was formatted following the instructions from the Sanger webpage (https://www.sanger.ac.uk/science/tools/sanger-imputation-service). This implied the removal (from the dataset to be imputed) of 152 tri-allelic markers, and 1608 insertions/deletions.

After imputation we tested the concordance of the imputed Variants in a subset of 143 samples that had been previously sequenced within GoNL. For this we selected only well imputed Variants (N=10,002,031), selected using the imputation quality score higher than 0.4 for Variants with a MAF>0.01, and higher than 0.8 for rare Variants (MAF<0.01). We used BCFtools v1.7 (http://samtools.github.io/bcftools/bcftools.html) and the command `filter -i`.

Overall, the minimum sample concordance was >98% (**Figure 22**, left), and only a couple of thousands of Variants had a low concordance (<50%; **Figure 22**, right), which could be explained by low sequence depth in GoNL. We observed that there was a tendency for the low MAF Variants to have lower concordance, but still, the great majority have a concordance near to 100% (**Figure 23**).

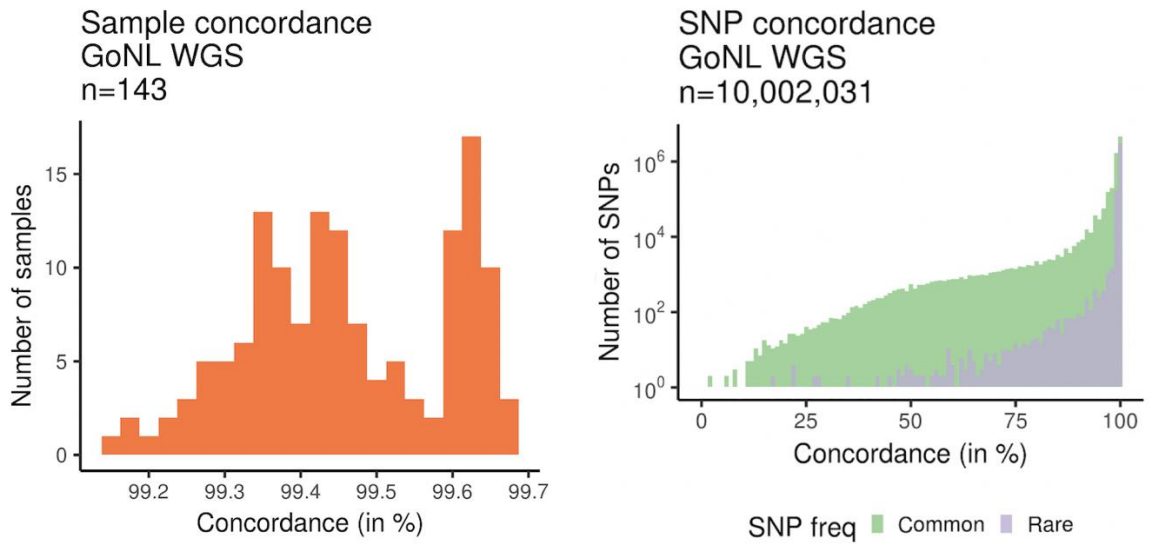

Figure 22. Concordance between imputed UGLI samples and sequenced GoNL samples. Left: sample concordance, right: SNP concordance.The bars on the left represent number of samples per category, the bars on the right indicate cumulative number of Variants.

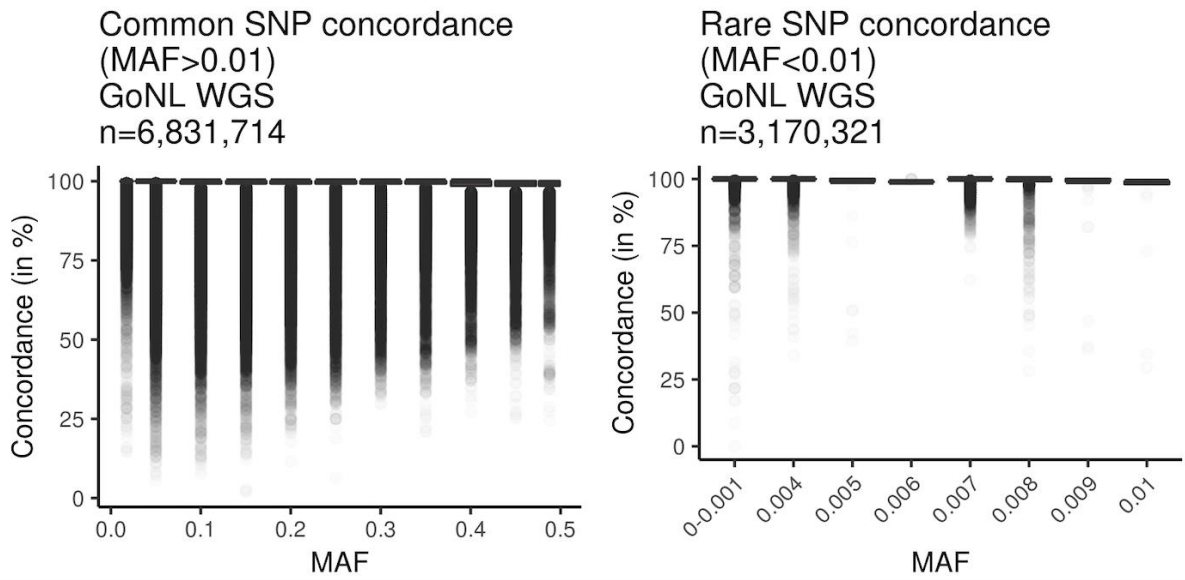

Figure 23. Concordance between imputed UGLI Variants and sequenced GoNL Variants. Left: Variants with MAF>0.01, right: Variants with MAF<0.01.

We used qctool v2.01 (https://www.well.ox.ac.uk/~gav/qctool_v2/index.html) commands `--vcf-genotype-field GP` to convert imputed files to bgen files. The imputed data set (all imputed markers, without filter for imputation quality) is available to researchers, together with the genotyping-only data set, where tri-allelic markers and insertions/deletions removed prior imputation can be found. To facilitate researchers, we provide log files with the information of these markers removed prior imputation, ethnicity information per samples, all removal steps, and a separate file with the first 20 PCs for population stratification.

Concluding remarks

We completed QC process of UGLI data, and constructed release 1 of genotyping and imputed data to be used for research. A log file with detailed per sample[2] and per SNP quality assessment is available to Lifelines.

Bibliography

[2] The log file for samples contains 120 additional samples (amounting to a total of 38,150 samples) that had to be re-genotyped and were not included in the release 1 for quality control. these will be processed in future releases

1. Shah TS, Liu JZ, Floyd JAB, Morris JA, Wirth N, Barrett JC, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics. 2012 Jun 15;28(12):1598–603.
2. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.
3. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

Authors:

Serena Sanna, Patrick Deleen, Raul Aguirre-Gamboa, Gerben van der Vries, Ilja Nolte and Esteban Lopera-Maya.

date: 27/11/2019