Genome Wide Associations Studies (abbreviation GWAS) have reproducibly identified thousands of loci, providing insight into underlying pathways of disease, in some cases with translational and clinical impact. Still, many questions remain about the genetic architecture of diseases. At Lifelines, we want to support research on genetics and its role in diseases and, therefore, DNA samples of 15,400 adult Lifelines participants where gathered to create a SNP database using a genome-wide genotyping array for all these samples. This genetic database of 15,400 Lifelines participants is called GWAS.

Basic information

Genome-wide genotype data based on the Illumina HumanCytoSNP-12 BeadChip v2 array are currently available for 15,400 participants. All GWAS participants are independent (no biological family relations), Caucasian-ancestry samples of adults which were collected at Baseline assessment second visit.

SNP array

The 12-sample HumanCytoSNP-12 BeadChip is a powerful, whole-genome scanning panel designed for efficient, high-throughput analysis of genetic and structural variations that are the most relevant to human disease. Many types and sizes of structural variation in the human genome that affect phenotypes can be detected with the HumanCytoSNP-12 BeadChip, including duplications, deletions, amplifications, copy-neutral LOH, and mosaicism. This BeadChip includes a complete panel of genome-wide tag SNPs and markers targeting all regions of known cytogenetic importance. It incorporates 200,000 SNPs which cover around 250 genomic regions commonly screened in cytogenetics laboratories, including subtelomeric regions, pericentromeric regions, sex chromosomes, and targeted coverage in around 400 additional disease-related genes.

Sample selection

To prevent false-positive association, non-Caucasian samples are excluded. Caucasian origin of participants was determined by:

- The LifeLines phenotype database (self-report)
- Outlier (IBS) analysis
- Population stratification (using Eigenstrat*)
Additionally, samples were selected using self-reported family relations. After cleaning of the data, samples were compared with each other to determine the relationship by genetic similarity. If a pair of samples was indicated as first degree relatives, the sample with the best genotyping quality will be included.

**Quality checks**

Quality controls of the data are based on SNP filtering on minor allele frequency (MAF) above 0.001, Hardy-Weinberg equilibrium (HWE) P-value $>1e^{-4}$, call rate of 0.95 using Plink, and principal component analysis (PCA) to check for population outliers. Sex chromosomes were used to check for sex mismatches, meaning that they were excluded when mismatched. SNPs located on sex chromosomes were not included in the data.

**Imputation**

SNP data obtained from the array were used to map against human reference genomes, i.e. the Genome of The Netherlands (GoNL) release 5 and the 1000 Genomes phase1 v3 reference panels, using Minimac version 2012.10.3. Before imputation, the genotypes were pre-phased using SHAPEIT2 and aligned to the reference panels using Genotype Harmonizer in order to resolve strand issues. Cleaned pedigree files and in- and output files for different imputation algorithms were created in PLINK binary format. Imputation analysis was performed using Beagle 3.1.0. The MOLGENIS compute imputation pipeline was used to generate and monitor job scripts on the distributed file system.

**SNP array in summary**

- **1000G imputed**: missing genotypes of participants are imputed (or predicted) based on 1000G reference panel (taken from worldwide population) and the available generated genotypes of the SNP array.
- **GoNL imputed**: missing genotypes of participants are imputed (or predicted) based on the GoNL reference panel (taken from the Dutch population) and the available generated genotypes of the SNP array.
- **Unimputed**: Genotypes determined on the basis of the SNP array, not imputed against genomic reference panels.

**SNP genotype data release**

The following files are available through the Lifelines workspace on the HPC (Linux environment):

- files with phenotype data
- files with genotyped and imputed data
- quality control files:
  - list of excluded samples
  - list of excluded SNPs
- PCA component file
Expansion of the genomic data - UGLI

Another additional assessment, that aims to genotype an even larger Lifelines cohort than GWAS, is UGLI. UGLI is the abbreviation for UMCG Genetics Lifelines Initiative. Approximately 38,000 participants are currently being genotyped using the Illumina global screening array (GSA) Beadchip-24 v1.0.

This array contains approximately 1,000,000 SNPs and combines multi-ethnic genome-wide content, curated clinical research variants, and quality control (QC) markers for precision medicine research 11. The UGLI cohort consists of Caucasian-ancestry samples of adults 18 years and older and approximately 3000 children aged 8-17 at Baseline assessment including biological family relations. All genotypes obtained from the array will be mapped against the global Haplotype Reference Consortium (HRC) panel to impute more SNPs 12.

UGLI data is currently being generated, analysed and quality checked. UGLI data will be made available to members of the UGLI consortium based on specific proposals approved by the UGLI steering committee and by Lifelines. Once the data is released, UGLI data access will be restricted to UGLI consortium members who are working at or are affiliated with the University Medical Center Groningen (UMCG) due to technical reasons. In the future the UGLI data will also be made available to non-UMCG researchers, though the timeline for this is unclear.

Gastrointestinal health research - microbiome analysis

There is increasing insight into the role of bacterial composition in the intestine and upon the occurrence of (chronic) diseases. DEEP 13 and DAG3 are two additional assessments in which data from Lifelines participants was gathered to assess the role of the gut microbiome in the occurrence of chronic diseases. Besides the faeces samples used to generate the microbiome data, in the DAG3 study, questionnaires were sent to participants to gather phenotypic data on GI health symptoms by means of Rome III criteria questionnaire 14 and the Bristol Stool Form Scale 15. A total of 9,500 participants were included in the DAG3 dataset, of which 9,300 are Dutch Caucasians and 700 are children aged 8-17 years. Approximately 1500 DEEP participants were included in the GWAS dataset, and roughly9000 DAG3 participants will be included in the UGLI dataset. The combination of fully genotyped Lifelines participants and microbiome data will allow researchers to study host-microbe interactions and will also allow studies to move from association to causality 16.

DEEP data, i.e. meta-genomic sequencing of Lifelines DEEP participants and 16S sequencing of stool samples from Lifelines DEEP participants can be requested at EGA. You can request gender and age of the Lifelines DEEP participants on this website as well. Additional phenotypic data can be requested by following the Lifelines application process on our website.

Currently DAG3 data is not available for release. Please contact us for updates on release dates.
References


