lifelines genetic data

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.

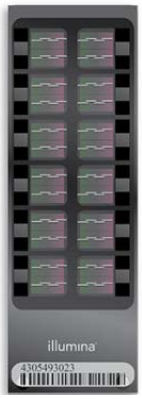
GWAS population general information	
Ratio male/female	41.8% / 58.2%
Average age*	47.8
Minimal Age*	18
Maximum Age*	89
Age category <18	N=0
Age category 18-64	N=13,890
Age category 65+	N=1,510

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[1] 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.



HumanCytoSNP-12 BeadChip SNP array. Source: www.illumina.com

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[2])

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 [3], 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 [4] and the 1000 Genomes phase1 v3 reference [5] panels, using Minimac version 2012.10.3 [9]. Before imputation, the genotypes were pre-phased using SHAPEIT2 [6] and aligned to the reference panels using Genotype Harmonizer [7] in order to resolve strand issues. Cleaned pedigree files and in- and output files for different imputation algorithms were created in PLINK [3] binary format. Imputation analysis was performed using Beagle 3.1.0 [8].The MOLGENIS compute imputation pipeline [10] 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].

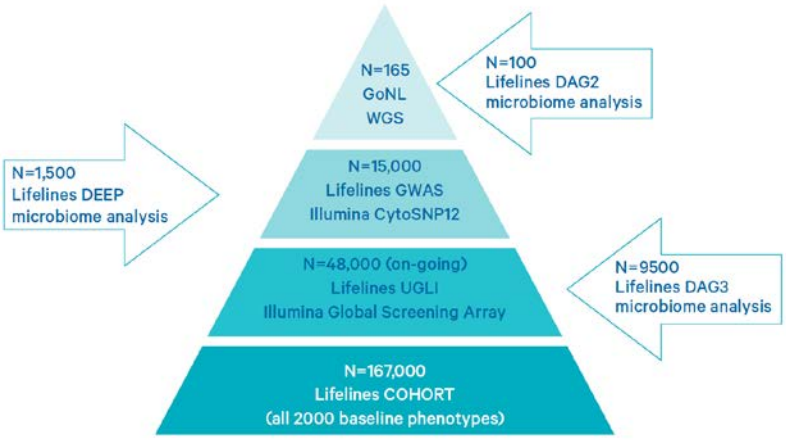


Figure 2: Overview of different genomic data using Lifelines participants.

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

1. https://www.illumina.com/products/by-type/clinical-research-products/human-cytosnp-12.html.

2. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847

3. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

4. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-825. doi:10.1038/ng.3021

5. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

6. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012;9(2):179-181. doi:10.1038/nmeth.1785

7. www.molgenis.org/systemsgenetics.

8. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*. 2008;84(2):210-223. doi:10.1016/j.ajhg.2009.01.005

9. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012;44(8):955-959. doi:10.1038/ng.2354

10. Byelas H, Byelas H, Dijkstra M, Neerincx P, Van Dijk F, Deelen AKP. Scaling bio-analyses from computational clusters to grids. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.415.9799. Accessed August 21, 2019.

11. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/infinium-commercial-gsa-data-sheet-370-2016-016.pdf.

12. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283. doi:10.1038/ng.3643

13. Tigchelaar EF, Zhernakova A, Dekens JAM, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. 2015;5(8):e006772. doi:10.1136/bmjopen-2014-006772

14. Longstreth GF, Thompson WG, Chey WD, Houghton LA, Mearin F, Spiller RC. Functional Bowel Disorders. *Gastroenterology*. 2006;130(5):1480-1491. doi:10.1053/j.gastro.2005.11.061

15. O'Donnell LJ, Virjee J, Heaton KW. Detection of pseudodiarrhoea by simple clinical assessment of intestinal transit rate. *BMJ*. 1990;300(6722):439-440. doi:10.1136/bmj.300.6722.439

16. Doestzada M, Vila AV, Zhernakova A, et al. Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein Cell*. 2018;9(5):432-445. doi:10.1007/s13238-018-0547-2