Family relations file

The file 'sec_family_relations_v3.1' is the most recent pedigree file available at Lifelines. We are still in the process of making this data available through the catalogue.

The file 'sec_family_relations_v3.1' is the updated version of 'sec_family_relations'. 'sec_family_relations' was only verified by using genetic data of UGLI 1 (GSA) and GWAS (CytoSNP). 'sec_family_relations_v3.1' has been verified by using genetic data from GWAS (CytoSNP), UGLI 1 (GSA), and UGLI 2 (Affymetrix).

The file 'sec_family_relations' was prepared by Serena Sanna who also was involved in the QC of the GSA data. The file 'sec_family_relations_v3.1' was prepared by Ilja Nolte who has also been preparing the Affymetrix data for release.

For more information, see the Lifelines wiki: http://wiki-lifelines.web.rug.nl/doku.php?id=family_relations

Information on preparation of 'sec_family_relations'

(Source: Serena Sanna, edited by Lifelines)

The basis of the family relations data within Lifelines consists of information from the municipal registries (Basisregistratie Personen (BRP)). Based on this information, researchers from the UGLI-consortium have optimized the family relations using recoded surname data, data from questionnaires, and genetic data.

UGLI first considered all blood-relationships declared by Lifelines participants (parent-child, sib-sib) and that allowed to link one Lifelines participant to another. UGLI constructed families based on this information, allowing consecutive relationships to be part of the same family (for example, merging grandparents to their children and grandchildren in one big family).

This information was not always complete - for example, in the case two Lifelines participants did not declare to be siblings and none of the parents were volunteers, the sib-sib relationship was not present in the database. Therefore, UGLI used additional information from follow up questionnaires such as "number of siblings", "year of birth of younger/older sibling" and coded (for privacy purposes) surnames, to identify additional sib-sib relationships.

In an earlier version, UGLI listed couples without children as two separate families. However, since research studies that are not confined to genetics are interested in a broad definition of family that also include partners, even if they haven't had children, UGLI also included the declared partner in the family in the 'sec_family_relations' file. Partner data of baseline was used.

Then, all gathered family information was coupled with genotyping data of GWAS (genotyped Lifelines population using the CytoSNP array) and UGLI 1 (genotyped Lifelines population using the Global Screening Array). All sib-sib relationships that were inferred through questionnaire data by UGLI (and could be checked by genetics) were correct, but coupling of genotyping data also identified additional sib-sib relationships and also cousin-cousin relationships. All these new relationships were confirmed using age and coded surname, and incorporated in the file 'sec_family_relations'.

Information on preparation of 'sec_family_relations_v3.1'

The file 'sec_family_relations_v3.1' is the updated version of 'sec_family_relations'. 'sec_family_relations_v3.1' has been verified by using genetic data from GWAS (CytoSNP), UGLI 1 (GSA), and UGLI 2 (Affymetrix).

Of note, partner data is based on baseline information.

Source of information below: UGLI2 (release 1.0) Quality Control Report, edited by Lifelines. See for more information the full UGLI2 (release 1.0) Quality Control Report.

A total of 147,833 known family relationships from the file 'sec_family_relations' were confirmed, while 18,689 (11.2%) relationships were flagged as errors. In addition 10,005 new relationships were found, of which 1,071 (10.7%) concerned first-degree relations (monozygotic twins/ duplicates, parent-offspring or full siblings) and 8934 concerned second-degree relations. UGLI analyzed any family relationship within families flagged as "error" that resulted in a genetically calculated first-degree or "unrelated" relationship (N=544, 0.3%; thus not analyzing genetically calculated 2nd, 3rd, and 4th relationships), as well as the 1,071 new first-degree relations between families.
If an error occurred in a family with only two genotyped family members (thus no genetic relation was found, while the file 'sec_family_relations' did state this relationship), 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) from UGLI2. In case of no sex discordance, both samples were excluded (N=104) from UGLI2. The relationship as mentioned in the 'sec_family_relations' file was kept in 'sec_family_relations_v3.1'. For each of the families with errors and that had more than two genotyped individuals (N=1,239), UGLI visualized the information in a pedigree plot, and 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. 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) 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.
The pedigree concordance analysis revealed that 144 errors occurred due to real monozygotic twins; 178 were full sibs that genetically turned out to be half sibs; 84 were corrections of parents within a family (i.e. a dummy was assigned, but the actual parent was present or full sibs that were thought to be half sibs); and 135 were identified as sample mix-ups. Of these sample mix-ups, 66 had enough genetic sharing with other Lifelines volunteers (i.e., relationships) to be reliably assigned to the correct individual (and family). The rest of the mixed-up samples (N=69) were excluded. Lists of samples swaps and samples to be excluded were created and with these files the Lifelines pedigree file, linkage files, and genotype PLINK fam files were corrected. The pedigree concordance analysis was repeated using these new files and verified that no additional sample mix-ups were present after this correction process. During the process, UGLI decided not (yet) to merge groups of families in which no other errors than duplicate samples between families occurred (i.e., a DNA sample which is identical to a DNA sample from another family), since this would only affect the Lifelines pedigree file and not whether UGLI2 samples should be swapped or excluded. Therefore there are still 308 groups of 2-6 families with first-degree relations that could be merged. UGLI plans to update this in the future.

Identifiers

The family relations file holds different types of PROJECT_PSEUDO_IDs.

- The long identifiers with numbers and letters combined, e.g. 0063btd8-0a2c-4370-9c65-90121e99ssd7 are the actual Lifelines identifiers that you will also see in the released data files.
- The S-numbers that are present in the family relations file are identifiers for mothers/fathers/children/partners who are not actually participating in Lifelines, but were derived based on the available information in the questionnaires. These are dummy identifiers that can be used to see if participants share a mother for instance. We do not have any other data available on persons with an S-number.
- Only for researchers that have made a catalogue selection: the P-numbers that are present in the family relations file are numbers of mothers/fathers/children/partners that do participate in Lifelines, but were not selected in your data order through the catalogue. For instance, if you selected only data on adults, the children participating in Lifelines will have a P-number in the family relations file.