**Tables and the content of each table V.0.0.1**

1. **Study-related meta data**

*This table will provide basic meta-data for each study plus time stamps for files. In the future number of columns will be extended to provide reach meta data. The tables below provide results from the separate analysis each study carried and those can differ from the overall entries for numbers, platforms, ancestry, etc. For example a study denoted as 'multiple' ancestry in this table should have separate results for European participants and participants from other ancestries in the table below. Similarly, here we will report the overall number of participants while, when possible, the number of participants contributing to each analyses will be reported in the tables bellow.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID and phenotype code  (e.g. PMID\_29212778\_EFO\_0000378 | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **phenotype code:**   * EFO * Ensembl gene, transcript, protein IDs * etc.   **This is a unique identifier for every study-phenotype pair.**  **EFO identifiers are primary and will be used whenever possible.** |
| study | unique identifier for the study | character | * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI; |
| cohort | names of cohort(s) that took part in the study | character | eg. UK Biobank, INTERVAL, etc. |
| study\_type | type of study | character | gwas, ewas, meta-gwas, wgs, pQTL, metQTL, MR, coloc, etc. |
| study\_year | year the study was published | integer | e.g. 2011 |
| study\_genotyping\_platform | type of genotyping platform | character | chip, GS, ES, multiple, etc. (including platform version if provided) |
| number\_of\_participants\_study | study-wide number of participants | integer | >0 |
| number\_of\_cases\_study | study-wide number of cases | integer | >0 |
| number\_of\_controls\_study | study-wide number of controls | integer | >0 |
| ancestry | study-wide ancestry | character | EUR, EAS, multiple, etc. |
| original\_build | Genome Reference Consortium Human Build (GRCh) of the original study | integer | 37,38,etc.  **in the database all studies should be mapped to 37 and later 38** |
| trait\_type | trait type | character | binary, quantitative |
| efo\_category | *different trait identifiers dependent on trait* | character | experimental factor ontology category |
| efo\_term | character | experimental factor ontology term |
| icd10 | character | International Statistical Classification of Diseases and Related Health Problems, version 10 |
| icd9 | character | International Statistical Classification of Diseases and Related Health Problems, version 9 |
| opcs4 | character | Classification of Medical Interventions and Procedures version 4 |
| opcs3 | character | Classification of Medical Interventions and Procedures version 3 |
| gene | character | Ensembl Gene ID |
| transcript | character | Ensembl Transcript ID |
| protein | character | Ensemble Protein ID |
| metabolite | code for the metabolite | character | CHEBI ID |
| source | which organ/tissue the data comes from | character | this is specifically for molecular data  uber-anatomy ontology code for organs and BrendaTissueOBO codes for cell types or serum |
| trait | trait name | character | human readable trait name (e.g. low-density lipoprotein) |
| trait\_abbreviation | colloquial abbreviation | character | e.g. LDL |
| file\_raw | name of raw file | character | the name of the file as downloaded from internet |
| time1\_raw | timestamp downloaded | date & time | the timestamp that goes with the database or date of first download if unavailable |
| time2\_raw | timestamp QCed | date & time | the timestamp of intermediate QC files |
| file\_processed | name of processed file | character | e.g. HDL\_study0\_chr(i).csv where i is chromosome identifier [1:25] |
| time1\_processed | timestamp of version in database | date & time | when multiple files per study exist the timestamp reflects the last file created |

**2.GWAS summary statistics**

*This table contains summary statistics estimates and data quality parameters that come from genome-wide association studies (GWAS) of binary and quantitative traits. The studies included here do not cover molecular traits unless they are covered by standard blood tests. In version V.0.0.1 the last column will always be NA as no imputation was done.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID and phenotype code  (e.g. PMID\_29212778\_EFO\_0000378  data used for building and testing purposes has "toy\_example" appended after the main identifier  **toy examples should not be used for analyses** | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **phenotype code:**   * EFO code or closest EFO   **This is a unique identifier for every study-phenotype pair.**  **EFO identifiers are primary and will be used whenever possible.** |
| efo\_term | phenotype code | character | this is the second part of the data\_set\_identifier; for this table efo\_term codes are primary and link to the meta-data table  relations with other codes can be found in meta-data table |
| chromosome | chromosome | integer | 1:25 (1:22,X,Y,MT) |
| position | position | integer | 1 to end of chromosome |
| reference\_allele | reference allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151 |
| alternative\_allele | alternative allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151   * **this is also the effect allele for the mapped data** |
| snp | SNP identifier | character | rsID (dbSNP 151)/other identifier from study (eg. chr:pos\_REF\_ALT) or NA if missing... |
| strand | which strand is the variant on | character | according to build 37: should be "+" |
| effect\_allele\_frequency | effect allele frequency | double | [0:1] |
| minor\_allele\_frequency | minor allele frequency | double | [0:0.5] |
| effect\_estimate | effect estimate | double | always beta; always ALT for build 37 |
| standard\_error | standard error | double | >0 |
| p | p-value | double | [0:1] |
| z | EFFECT/SE | double |  |
| genotype\_imputation\_score | genotype imputation score | double | [0:1];  INFO from IMPUTE or R2 from MACH |
| direction | direction of effect | character | '+',''-' |
| number\_of\_participants | number of individuals in the particular analysis | integer | total for case-control/ analysed for quantitative traits |
| number\_of\_cases | number of cases in the particular analysis | integer | NA for quantitative traits |
| number\_of\_controls | number of controls in the particular analysis | integer | NA for quantitative traits |
| hetisq | heterogeneity I squared | double | has values for meta analyses only |
| hetdf | heterogeneity degrees of freedom | integer | has values for meta analyses only |
| hetpval | heterogeneity p-value | double | has values for meta analyses only [0:1] |
| hweq | Hardy-Weinberg equilibrium | double | [0:1] |
| original\_effect\_allele | original effect allele | character | effect allele in study |
| original\_other\_allele | original other allele | character | other allele in study |
| original\_strand | original strand | character | according to study: could be "+" or "-" |
| original\_direction | original direction | character | according to study: '+',''-' or multiple pluses/minuses for meta-analyses |
| original\_effect\_allele\_frequency | original allele frequncy | double | [0:1] |
| statistics\_imputation\_score | statistics imputation score | double | imputation score for missing summary statistics; [0:1] |

**3.eQTL summary statistics**

*This table contains summary statistics estimates and data quality parameters that come from gene expression quantitative trait loci (eQTL) mapping studies. In addition to summary statistics it contains gene, transcript and tissues information. At first instance the table will include only cis-eQTLs. Later other types of eQTLs will be added.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID + gene  + source | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **gene:**Ensembl Gene ID  **source:** uber-anatomy ontology code for organs and BrendaTissueOBO codes for cell types or serum |
| gene | Ensembl gene ID | character | as in data\_set\_identifier |
| transcript | Ensembl transcript ID | character |  |
| source | which organ/tissue the data comes from | character | as in data\_set\_identifier |

|  |  |  |  |
| --- | --- | --- | --- |
| chromosome | chromosome | integer | 1:25 (1:22,X,Y,MT) |
| position | position | integer | 1 to end of chromosome |
| reference\_allele | reference allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151 |
| alternative\_allele | alternative allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151   * **this is also the effect allele for the mapped data** |
| snp | SNP identifier | character | rsID (dbSNP 151)/other identifier from study (eg. chr:pos\_REF\_ALT) or NA if missing... |
| strand | which strand is the variant on | character | according to build 37: should be "+" |
| effect\_allele\_frequency | effect allele frequency | double | [0:1] |
| minor\_allele\_frequency | minor allele frequency | double | [0:0.5] |
| effect\_estimate | effect estimate | double | always beta; always ALT for build 37 |
| standard\_error | standard error | double | >0 |
| z | EFFECT/SE | double |  |
| p | p-value | double | [0:1] |
| genotype\_imputation\_score | genotype imputation score | double | [0:1];INFO from IMPUTE or R2 from MACH |
| direction | direction of effect | character | '+',''-' |
| number\_of\_participants | number of individuals in the particular analysis | integer | total for case-control/ analysed for quantitative traits |
| number\_of\_cases | number of cases in the particular analysis | integer | NA for quantitative traits |
| number\_of\_controls | number of controls in the particular analysis | integer | NA for quantitative traits |
| hetisq | heterogeneity I squared | double | has values for meta analyses only |
| hetdf | heterogeneity degrees of freedom | integer | has values for meta analyses only |
| hetpval | heterogeneity p-value | double | has values for meta analyses only [0:1] |
| hweq | Hardy-Weinberg equilibrium | double | [0:1] |

|  |  |  |  |
| --- | --- | --- | --- |
| original\_effect\_allele | original effect allele | character | effect allele in study |
| original\_other\_allele | original other allele | character | other allele in study |
| original\_strand | original strand | character | according to study: could be "+" or "-" |
| original\_direction | original direction | character | according to study: '+',''-' or multiple pluses/minuses for meta-analyses |
| original\_effect\_allele\_frequency | original allele frequency | double | [0:1] |
| statistics\_imputation\_score | statistics imputation score | double | imputation score for missing summary statistics; [0:1] |

**4.pQTL summary statisitcs table**

*This table contains summary statistics estimates and data quality parameters that come from protein expression quantitative trait loci (pQTL) mapping studies. In addition to summary statistics it contains protein and tissues information.  Both cis and trans effects are included.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID + protein ID+ source code | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **protein:**Ensembl Gene ID  **source:** uber-anatomy ontology code for organs and BrendaTissueOBO codes for cell types or serum |
| gene | Ensembl gene ID | character |  |
| protein | Ensembl protein ID | character | as in data\_set\_identifier |
| source | which organ/tissue the data comes from | character | as in data\_set\_identifier |

|  |  |  |  |
| --- | --- | --- | --- |
| chromosome | chromosome | integer | 1:25 (1:22,X,Y,MT) |
| position | position | integer | 1 to end of chromosome |
| reference\_allele | reference allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151 |
| alternative\_allele | alternative allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151   * **this is also the effect allele for the mapped data** |
| snp | SNP identifier | character | rsID (dbSNP 151)/other identifier from study (eg. chr:pos\_REF\_ALT) or NA if missing... |
| strand | which strand is the variant on | character | according to build 37: should be "+" |
| effect\_allele\_frequency | effect allele frequency | double | [0:1] |
| minor\_allele\_frequency | minor allele frequency | double | [0:0.5] |
| effect\_estimate | effect estimate | double | always beta; always ALT for build 37 |
| standard\_error | standard error | double | >0 |
| p | p-value | double | [0:1] |
| z | EFFECT/SE | double |  |
| genotype\_imputation\_score | genotype imputation score | double | [0:1];INFO from IMPUTE or R2 from MACH |
| direction | direction of effect | character | '+',''-' |
| number\_of\_participants | number of individuals in the particular analysis | integer | total for case-control/ analysed for quantitative traits |
| number\_of\_cases | number of cases in the particular analysis | integer | NA for quantitative traits |
| number\_of\_controls | number of controls in the particular analysis | integer | NA for quantitative traits |
| hetisq | heterogeneity I squared | double | has values for meta analyses only |
| hetdf | heterogeneity degrees of freedom | integer | has values for meta analyses only |
| hetpval | heterogeneity p-value | double | has values for meta analyses only [0:1] |
| hweq | Hardy-Weinberg equilibrium | double | [0:1] |

|  |  |  |  |
| --- | --- | --- | --- |
| original\_effect\_allele | original effect allele | character | effect allele in study |
| original\_other\_allele | original other allele | character | other allele in study |
| original\_strand | original strand | character | according to study: could be "+" or "-" |
| original\_direction | original direction | character | according to study: '+',''-' or multiple pluses/minuses for meta-analyses |
| original\_effect\_allele\_frequency | original allele frequency | double | [0:1] |
| statistics\_imputation\_score | statistics imputation score | double | imputation score for missing summary statistics; [0:1] |

|  |  |  |  |
| --- | --- | --- | --- |
| name | protein name as is on the platform | character | SOMAscan/OLINK or other name |

**5.mQTL summary statisitce table**

*This table contains summary statistics estimates and data quality parameters that come from metabolite level quantitative trait loci (mQTL) mapping studies.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID + protein ID+ source code | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **metabolite:**CHEBI ID  **source:** uber-anatomy ontology code for organs and BrendaTissueOBO codes for cell types or serum |
| metabolite | CHEBI ID | character | as in data\_set\_identifier |
| source | which organ/tissue the data comes from | character | as in data\_set\_identifier |

|  |  |  |  |
| --- | --- | --- | --- |
| chromosome | chromosome | integer | 1:25 (1:22,X,Y,MT) |
| position | position | integer | 1 to end of chromosome |
| reference\_allele | reference allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151 |
| alternative\_allele | alternative allele | character | according to Genome Reference Consortium Human Build 37 (GRCh37) & dbSNP 151   * **this is also the effect allele for the mapped data** |
| snp | SNP identifier | character | rsID (dbSNP 151)/other identifier from study (eg. chr:pos\_REF\_ALT) or NA if missing... |
| strand | which strand is the variant on | character | according to build 37: should be "+" |
| effect\_allele\_frequency | effect allele frequency | double | [0:1] |
| minor\_allele\_frequency | minor allele frequency | double | [0:0.5] |
| effect\_estimate | effect estimate | double | always beta; always ALT for build 37 |
| standard\_error | standard error | double | >0 |
| z | EFFECT/SE | double |  |
| p | p-value |  | [0:1] |
| genotype\_imputation\_score | genotype imputation score | double | [0:1];INFO from IMPUTE or R2 from MACH |
| direction | direction of effect | character | '+',''-' |
| number\_of\_participants | number of individuals in the particular analysis | integer | total for case-control/ analysed for quantitative traits |
| number\_of\_cases | number of cases in the particular analysis | integer | NA for quantitative traits |
| number\_of\_controls | number of controls in the particular analysis | integer | NA for quantitative traits |
| hetisq | heterogeneity I squared | double | has values for meta analyses only |
| hetdf | heterogeneity degrees of freedom | integer | has values for meta analyses only |
| hetpval | heterogeneity p-value | double | has values for meta analyses only [0:1] |
| hweq | Hardy-Weinberg equilibrium | double | [0:1] |
| original\_effect\_allele | original effect allele | character | effect allele in study |
| original\_other\_allele | original other allele | character | other allele in study |
| original\_strand | original strand | character | according to study: could be "+" or "-" |
| original\_direction | original direction | character | according to study: '+',''-' or multiple pluses/minuses for meta-analyses |
| original\_effect\_allele\_frequency | original allele frequncy | double | [0:1] |
| statistics\_imputation\_score | statistics imputation score | double | imputation score for missing summary statistics; [0:1] |

|  |  |  |  |
| --- | --- | --- | --- |
| name | name as on platform | character | e.g. Metabolon/NMR identifier |

**6.EWAS summary statistics table**

*This table contains summary statistics estimates and data quality parameters that come from epigenome-wide association studies (EWAS).*

*The table will primarily host methylation results but other types of epigenetic marks will be added when possible.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| data\_set\_identifier | study ID + probe name + source code | character | **study ID:**   * PMID * bioRxiv DOI * internal code * provisional codes if unpublished and not in bioRxiv DOI;   **metabolite:**CHEBI ID  **source:** uber-anatomy ontology code for organs and BrendaTissueOBO codes for cell types or serum |
| probe\_name | methylation probe name | character |  |
| chromosome | chromosome | integer |  |
| position | position | integer |  |
| reference allele | reference allele | character |  |
| alternative allele | alternative allele | character | == effect allele |
| snp | rsID | character |  |
| strand | which strand is the variant on | character | according to build 37: should be "+" |
| probe\_chromosome | methylation probe chromosome | integer |  |
| probe\_position | the centre of the methylation probe position | integer |  |
| type | type of metQTL | character | cys or trans |
| effect\_allele\_frequency | effect allele frequency | double | [0:1] |
| minor\_allele\_frequency | minor\_allele\_frequency | double | [0:0.5] |
| effect estimate | effect estimate | double |  |
| standard error | standard error | double | >0 |
| z | EFFECT/SE | double |  |
| p | p-value | double | [0:1] |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| genotype\_imputation\_score | | genotype imputation score | | | double | [0:1];INFO from IMPUTE or R2 from MACH |
| direction | | direction of effect | | | character | '+',''-' |
| number\_of\_participants | | number of individuals in the particular analysis | | | integer | total for case-control/ analysed for quantitative traits |
| number\_of\_cases | | number of cases in the particular analysis | | | integer | NA for quantitative traits |
| number\_of\_controls | | number of controls in the particular analysis | | | integer | NA for quantitative traits |
| hetisq | | heterogeneity I squared | | | double | has values for meta analyses only |
| hetdf | | heterogeneity degrees of freedom | | | integer | has values for meta analyses only |
| hetpval | | heterogeneity p-value | | | double | has values for meta analyses only [0:1] |
| hweq | | Hardy-Weinberg equilibrium | | | double | [0:1] |
| original\_effect\_allele | | original effect allele | | | character | effect allele in study |
| original\_other\_allele | | original other allele | | | character | other allele in study |
| original\_strand | | original strand | | | character | according to study: could be "+" or "-" |
| original\_direction | | original direction | | | character | according to study: '+',''-' or multiple pluses/minuses for meta-analyses |
| original\_effect\_allele\_frequency | | original allele frequncy | | | double | [0:1] |
| statistics\_imputation\_score | | statistics imputation score | | | double | imputation score for missing summary statistics; [0:1] |
| gene | Ensembl gene ID | | character |  | | |
| fdr | falce discovery rate q-value | | double | [0:1] | | |
| name | study ID and phenotype code | | character | gene name according to study (e.g. HGNCName) | | |

**7.Allele\_freq\_reference\_cohorts**

(comment REF+ALT for a population = 1)

*This table will provides allele frequencies for a variety of reference cohorts that will be used for quality control and comparison purposes. In this version we start with 1000 Genomes data.*

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| chromosome | chromosome | integer | chr, pos and allele form the unique identifier to the allele\_freq\_ref table |
| position | position | integer | chr, pos and allele form the unique identifier to the allele\_freq\_ref table |
| allele | sequence of the alt allele | varchar(100) | chr, pos and allele form the unique identifier to the allele\_freq\_ref table |
| cohort | name of the cohort/dataset | varchar | e.g. 1000 Genomes |
| ethnicity | ethnicity code | varchar | refers to the ethnicity\_codes table |
| genotyping\_method | a code for the genotyping method | varchar | [GS, ES, chip,dbSNP151] |
| alternative\_allele\_frequency | alternative allele frequency | double |  |
| minor\_allele | T if alt\_frequency < 0.5 | boolean | Need a DB function which will set this to true if the alt\_frequency is < 0.5 and false if it is >=0.5 |

Table allele\_freq\_ref as a\_ref {  
   chr int [pk]  
   pos int [pk]  
   allele varchar(100) [pk] // allele in build 37  
   type varchar(3) // ref, alt  
 }  
  
 Table allele\_freq\_cohort as a\_coh {  
   allele\_freq\_cohort\_id int [pk, increment] // auto-increment  
   chr int  
   pos int  
   allele varchar(100)  
   cohort varchar // name of the population   
   ethnicity varchar // ethnicity code  
   genotyping\_method varchar // e.g. GS, ES, chip, dbSNP151  
   alt\_frequency double // 0:1  
   minor\_allele boolean // make T if alt\_frequency < 0.5 {This should be a calculated by a function in the DB}  
 }  
  
 Table ethnicity\_codes as ec {  
   ethnicity varchar [pk]  
   description varchar  
 }

Ref: a\_ref.(chr, pos, allele) < a\_coh.(chr, pos, allele)  
Ref: a\_coh.ethnicity - ec.ethnicity

**Table 8: Allele MAP**

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| chromosome | chromosome | integer [pk] | 1:25 (1:22,X,Y,MT) |
| position | position | integer [pk] |  |
| allele | sequence of the alternative allele | varchar(100) [pk] | allele in build 37 (1-100 length) |
| type | type of allele (ref or alt) | varchar(3) | [ref, alt] |

Q - do we need a SNP id in the above table?

**Table 9: ethnicity\_codes**

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| ETHNICITY | ethnicity code | varchar [pk] | a standard code or abbreviation |
| DESCRIPTION | a description of the ethnicity | varchar |  |

**Table10:Variant\_annotation**

|  |  |  |  |
| --- | --- | --- | --- |
| Name of column | Meaning | Type | Notes |
| chromosome | chromosome | integer |  |
| position | position | integer |  |
| allele | alternative allele | character |  |
| vep\_sequance\_change | sequence change | character | e.g. a "-" if deletion |
| gene | Ensembl gene ID | character |  |
| feature | Ensembl stable ID of feature | character |  |
| feature\_type | Type of feature | character | Currently one of Transcript, RegulatoryFeature, MotifFeature. |
| consequence | consequence for each transcript | character | <https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences> |
| cdna\_position | relative position of base pair in cDNA sequence | integer |  |
| cds\_position | relative position of base pair in coding sequence | integer |  |
| protein\_position | relative position of amino acid in protein | integer |  |
| amino\_acids | amino acid change | character | only given if the variant affects the protein-coding sequence |
| codons | the alternative codons | character | variant base in upper case |
| existing\_variation | additional annotation from VEP subroutines | boolean |  |
| distance | integer |  |
| strand | integer |  |
| sift | character |  |
| polyphen | character |  |
| motif\_name | character |  |
| motif\_position | integer |  |
| high\_inf\_position | character |  |
| motif\_score\_changes | double |  |

**Connections between tables**

Table 1 (Meta data) to summary statistics tables (2-6) by IDENTIFIER

Tables 2-6 between each other  CHR, POS, REF, ALT

Table 7-10 see figure above

**Notes**

1) In current version we have chromosomes 1-22, chromosomes X,Y,MT will be added in later

2) FDR for molecular traits in all tables?

3) Full version control to be implemented in V.0.0.2

4) At later stage it will be good to add model codes

5) In the future, there will data-base specific there will be variant identifiers

6) We also want to ad metagenomics results at some point in time. That would be a separate table...