Variant Annotations Help File

The set of files comprised of **var_pheno_ann.tsv**, **var_drug_ann.tsv**, **var_fa_ann.tsv** and **study_parameters.tsv** contain PharmGKB's variant annotations and associated information. Variant annotations are separated into 3 files because PharmGKB has 3 different standardized templates to capture information depending on the specifics of the association. Each file has slightly different fields, though some fields are common to all 3. Please refer to the PharmGKB website for more information about variant annotations, and how they are used to create clinical annotations based on score.

It is important to understand that the PharmGKB curators routinely review several high profile journals for articles to curate. There may be more literature in the public domain to support or contradict an association that is not in the PharmGKB database. PharmGKB does its best to manually curate high profile literature but does not contain curated literature from every domain-based journal, or all of PubMed. PharmGKB reviews evidence from curated literature in non-regular intervals and re-evaluates the evidence strength for each association as more literature becomes available.

Description of Files:

- **var_pheno_ann.tsv**: Contains associations in which the variant affects a phenotype, with or without drug information.
- **var_drug_ann.tsv**: Contains associations in which the variant affects a drug dose, response, metabolism, etc.
- **var_fa_ann.tsv**: Contains *in vitro* and functional analysis-type associations.
- **study_parameters.tsv**: Contains information about the study population size, biogeographical group and statistics for the variant annotations; this file is cross-referenced against the 3 variant annotation files.
- **LICENSE.txt**: The PharmGKB license for using PharmGKB data, including variant and clinical annotations.
- **CREATED_xxxx-xx-xx.txt**: This file indicates the date that all files in this group were created from the database.
- **README.pdf** file: This document.

A description of the fields in each file follows.

var_pheno_ann.tsv:

- ***Variant Annotation ID***: Unique ID number for each variant/drug annotation.
- ***Variant/Haplotypes***: dbSNP rsID or haplotype(s) involved in the association. In some cases, an association is based on a gene phenotype group such as "poor metabolizers" or "intermediate activity". In these cases, the gene phenotype is found in this field.
- ***Gene***: HGNC symbol for the gene involved in the association. Typically the variants will be *within* the gene boundaries, but occasionally this will not be true. E.g. the variant in the annotation may be upstream of the gene but is reported to affect the gene's expression or otherwise associated with the gene.
- ***Drug(s)***: The drug(s) involved in the association. If there is more than one drug listed, the association may apply to each drug individually or the combination of the drugs together. The field "Multiple drugs And/or" will designate "or" - meaning that it applies to each drug - or "and" - meaning that the association is for the combination.
- ***PMID***: PubMed identifier for the article supporting the annotation.
- ***Phenotype Category***: Options are "efficacy", "toxicity", "dosage", "metabolism/PK", "PD", "other".
- ***Significance***: The significance of the association as stated by the author; options are [yes, no, not stated].

- ***Notes***: Free text field for notes added by the curator.
- ***Sentence***: The structured annotation sentence generated by the variant annotation tool based on the information entered by the curator.
- ***Alleles***: The basis for comparison in the annotation. In this field, there may be a variant, one or more haplotypes grouped together, one or more genotypes grouped together or one or more diplotypes grouped together. If there is a gene phenotype in the "Variant/Haplotypes" field (described above), this field will be blank.
- ***Specialty Population***: Any special populations this annotation is relevant to (e.g. pediatric).
- The following columns are parts of the structured sentence in the "Sentence" field above. Note that not all fields are required.
    - ***Metabolizer types***: This field contains the gene phenotype group, if applicable. I.e., if the association is based on a gene phenotype group such as "poor metabolizers", or if the association is based on individual genotypes that were combined into a gene phenotype group.
    - ***isPlural***: This field maintains grammar in the sentence. Options are "is" and "are".
    - ***Is/Is Not associated***: This field indicates whether or not an association was found. Options are "associated with" or "not associated with".
    - ***Direction of effect***: This field is a descriptor of the "Side effect/efficacy/other" field. Options are "increased" or "decreased".
    - ***Side effect/efficacy/other***: This field describes the "Phenotype" of the association. Options are "likelihood of", "risk of", "severity of" or "age at onset of".
    - ***Phenotype***: This field is the resulting phenotype of the association. It can be a "disease" from the PharmGKB standardized vocabulary, a "side effect", an "efficacy" term or "other". This field can be free text or an existing term in PharmGKB; it is not standardized. There can also be multiple phenotypes entered, in this field, separated by commas ",".
    - ***Multiple phenotypes And/or***: If there is more than one entry in the "Phenotype" field, this field specifies if the phenotypes should be separated by "and" or "or". Options are "and" or "or".
    - ***When treated with/exposed to/when assayed with***: This field contains words that connect the first part of the sentence with the next part. There are many options including, but not limited to, "when treated with", "when exposed to" or "when assayed with".
    - ***Multiple drugs And/or***: If there is more than one drug in the "Drug(s)" field, "or" designates that the association applies to each drug individually, or "and" designates the association applies to the combination of the drugs together. Options are "and" or "or".
    - ***Population types***: This field indicates the type of population in which the association was studied. There are multiple options including, but not limited to, "in healthy individuals", "in children", "in women". The studied population can be further described in the "Population Phenotypes or diseases" field, in which case the "Population types" field will in with the word "with", such as "in children with".
    - ***Population Phenotypes or diseases***: This field further describes the studied population entered in the "Population types" field by indicating if the population has a particular disease or phenotype such as "diabetes" or "lung transplantation".
    - ***Multiple phenotypes or diseases And/or***: If there is more than one entry in the "Population Phenotypes or diseases" field, this field specifies if the phenotypes should be separated by "and" or "or". Options are "and" or "or".
    - ***Comparison Allele(s) or Genotype(s)***: This field indicates if the study authors directly compared the allele(s) or genotype(s) in the "Variant/Haplotypes" field against other allele(s)/haplotype(s)/genotype(s).
    - ***Comparison Metabolizer types***: This field indicates if the study authors directly compared the gene phenotype group in the "Metabolizer types" field against other gene phenotype groups.

var_drug_ann.tsv:

- ***Variant Annotation ID***: Unique ID number for each variant/drug annotation.
- ***Variant/Haplotypes***: dbSNP rsID or haplotype(s) involved in the association. In some cases, an association is based on a gene phenotype group such as "poor metabolizers" or "intermediate activity". In these cases, the gene phenotype is found in this field.
- ***Gene***: HGNC symbol for the gene involved in the association. Typically the variants will be *within* the gene boundaries, but occasionally this will not be true. E.g. the variant in the annotation may be upstream of the gene but is reported to affect the gene's expression or otherwise associated with the gene.
- ***Drug(s)***: The drug(s) involved in the association. Drugs are standardized objects in PharmGKB, typically based on the generic drug name. If there is more than one drug listed, the association may apply to each drug individually or the combination of the drugs together. The field "Multiple drugs And/or" will designate "or" - meaning that it applies to each drug - or "and" - meaning that the association is for the combination.
- ***PMID***: PubMed identifier for the article supporting the annotation.
- ***Phenotype Category***: Options are "efficacy", "toxicity", "dosage", "metabolism/PK", "PD", "other".
- ***Significance***: The significance of the association as stated by the author; options are [yes, no, not stated].
- ***Notes***: Free text field for notes added by the curator.
- ***Sentence***: The structured annotation sentence generated by the variant annotation tool based on the information entered by the curator.
- ***Alleles***: The basis for comparison in the annotation. In this field, there may be a variant, one or more haplotypes grouped together, one or more genotypes grouped together or one or more diplotypes grouped together. If there is a gene phenotype in the "Variant/Haplotypes" field (described above), this field will be blank.
- ***Specialty Population***: Any special populations this annotation is relevant to (e.g. pediatric).
- The following columns are parts of the structured sentence in the "Sentence" field above. Note that not all fields are required.
    - ***Metabolizer types***: This field contains the gene phenotype group, if applicable. I.e., if the association is based on a gene phenotype group such as "poor metabolizers", or if the association is based on individual genotypes that were combined into a gene phenotype group.
    - ***isPlural***: This field maintains grammar in the sentence. Options are "is" and "are".
    - ***Is/Is Not associated***: This field indicates whether or not an association was found. Options are "associated with" or "not associated with".
    - ***Direction of effect***: This field is a descriptor of the "PD/PK terms" field. Options are "increased" or "decreased".
    - ***PD/PK terms***: This field contains the pharmacodynamic or pharmacokinetic phenotype that was measured in the association. This field also serves to connect the variant and drug together in the standardized sentence. There are many options including, but not limited to, "concentration of", "metabolism of" or "response to".
    - ***Multiple drugs And/or***: If there is more than one drug in the "Drug(s)" field, "or" designates that the association applies to each drug individually, or "and" designates the association applies to the combination of the drugs together. Options are "and" or "or".
    - ***Population types***: This field indicates the type of population in which the association was studied. There are multiple options including, but not limited to, "in healthy individuals", "in children", "in women". The studied population can be further described in the "Population Phenotypes or diseases" field, in which case the "Population types" field will in with the word "with", such as "in children with".
    - ***Population Phenotypes or diseases***: This field further describes the studied population entered in the

"Population types" field by indicating if the population has a particular disease or phenotype such as "diabetes" or "lung transplantation".
    - ***Multiple phenotypes or diseases And/or***: If there is more than one entry in the "Population Phenotypes or diseases" field, this field specifies if the phenotypes should be separated by "and" or "or". Options are "and" or "or".
    - ***Comparison Allele(s) or Genotype(s)***: This field indicates if the study authors directly compared the allele(s) or genotype(s) in the "Variant/Haplotypes" field against other allele(s)/haplotype(s)/genotype(s).
    - ***Comparison Metabolizer types***: This field indicates if the study authors directly compared the gene phenotype group in the "Metabolizer types" field against other gene phenotype groups.

var_fa_ann.tsv:

- ***Variant Annotation ID***: Unique ID number for each variant/drug annotation.
- ***Variant/Haplotypes***: dbSNP rsID or haplotype(s) involved in the association. In some cases, an association is based on a gene phenotype group such as "poor metabolizers" or "intermediate activity". In these cases, the gene phenotype is found in this field.
- ***Gene***: HGNC symbol for the gene involved in the association. Typically the variants will be *within* the gene boundaries, but occasionally this will not be true. E.g. the variant in the annotation may be upstream of the gene but is reported to affect the gene's expression or otherwise associated with the gene.
- ***Drug(s)***: The drug(s) involved in the association. If there is more than one drug listed, the association may apply to each drug individually or the combination of the drugs together. The field "Multiple drugs And/or" will designate "or" - meaning that it applies to each drug - or "and" - meaning that the association is for the combination.
- ***PMID***: PubMed identifier for the article supporting the annotation.
- ***Phenotype Category***: Options are "efficacy", "toxicity", "dosage", "metabolism/PK", "PD", "other".
- ***Significance***: The significance of the association as stated by the author; options are [yes, no, not stated].
- ***Notes***: Free text field for notes added by the curator.
- ***Sentence***: The structured annotation sentence generated by the variant annotation tool based on the information entered by the curator.
- ***Alleles***: The basis for comparison in the annotation. In this field, there may be a variant, one or more haplotypes grouped together, one or more genotypes grouped together or one or more diplotypes grouped together. If there is a gene phenotype in the "Variant/Haplotypes" field (described above), this field will be blank.
- ***Specialty Population***: Any special populations this annotation is relevant to (e.g. pediatric).
- ***Assay Type***: Information about the type of assay performed.
- The following columns are parts of the structured sentence in the "Sentence" field above. Note that not all fields are required.
    - ***Metabolizer types***: This field contains the gene phenotype group, if applicable. I.e., if the association is based on a gene phenotype group such as "poor metabolizers", or if the association is based on individual genotypes that were combined into a gene phenotype group.
    - ***isPlural***: This field maintains grammar in the sentence. Options are "is" and "are".
    - ***Is/Is Not associated***: This field indicates whether or not an association was found. Options are "associated with" or "not associated with".
    - ***Direction of effect***: This field is a descriptor of the "Functional terms" field. Options are "increased" or "decreased".
    - ***Functional terms***: This field contains the functional phenotype that was measured in the association. This field also serves to connect the variant and drug together in the standardized sentence. There are many

options including, but not limited to, "activity of", "expression of" or "inhibition of".
    - ***Gene/gene product***: This field contains HGNC gene symbol for the gene (product) being measured by the functional assay. This field is not required, but if an entry is present, it should match the gene in the "Gene" field.
    - ***When treated with/exposed to/when assayed with***: This field contains words that connect the first part of the sentence with the next part. There are many options including, but not limited to, "when assayed with", "when exposed to" or "due to".
    - ***Multiple drugs And/or***: If there is more than one drug in the "Drug(s)" field, "or" designates that the association applies to each drug individually, or "and" designates the association applies to the combination of the drugs together. Options are "and" or "or".
    - ***Cell type:***: This field contains the type of cell in which the assay was conducted. This field is free text.
    - ***Comparison Allele(s) or Genotype(s)***: This field indicates if the study authors directly compared the allele(s) or genotype(s) in the "Variant/Haplotypes" field against other allele(s)/haplotype(s)/genotype(s).
    - ***Comparison Metabolizer types***: This field indicates if the study authors directly compared the gene phenotype group in the "Metabolizer types" field against other gene phenotype groups.

study_parameters.tsv:

- ***Study Parameters ID***: Unique ID number for each "Study Parameters" entry.
- ***Variant Annotation ID***: ID number of the associated Variant Annotation. The ID number in this field will match with a "Variant Annotation ID" in one of the three variant annotation files.
- ***Study Type***: The type of study reported in the paper; options are "cohort", "case/control", "case series", "cross-sectional", "clinical trial", "meta-analysis", "GWAS", "replication", "prospective", "retrospective", "linkage", "trios".
- ***Study Cases***: The number of cases in the paper.
- ***Study Controls***: The number of controls used in the association analysis in the paper.
- ***Characteristics***: Free text entered by the curator to record details such as gender, disease, age group or other distinguishing characteristics about the group studied.
- ***Characteristics Type***: The standardized term used to describe the "Characteristics" field; options are "disease", "drug", "age group", "gender", "study cohort".
- ***Frequency in Cases***: Allele frequency in the cases *if clearly reported in the paper*.
- ***Allele of Frequency in Cases***: The allele the "Frequency in Cases" refers to.
- ***Frequency in Controls***: Allele frequency in the controls *if clearly reported in the paper*.
- ***Allele of Frequency in Controls***: The allele the "Frequency in Controls" refers to.
- ***P Value***: The p-value and the operator (=, <, etc...) that is reported in the paper.
- ***Ratio Stat Type***: The type of statistic reported in the paper; options are "OR" (odds ratio), "RR" (relative risk) or "HR" (hazard ratio).
- ***Ratio Stat:*** The number associated with the "Ratio Stat Type" reported in the paper.
- ***Confidence Interval Start***: The start of the confidence interval for the "Ratio Stat".
- ***Confidence Interval Stop***: The end of the confidence interval for the "Ratio Stat".
- ***Biogeographical Groups***: The population groups from PMID:30506572 (more information found at https://www.pharmgkb.org/page/biogeographicalGroups). Options are "African America/Afro-Caribbean", "American", "Central/South Asian", "East Asian", "European", "Latino", "Near Eastern", "Oceanian", "Sub-Saharan African", "Unknown", "Multiple Groups". The "Multiple Groups" option indicates that the study group was comprised of people from more than one biogeographical population group. In many cases, details are included

when the "Multiple Groups" option is used.