Standards Overview

PharmVar uses a number of conventions for storing and displaying allelic data consistently. This includes reliance on public standards and data sources wherever possible.

Reference Sequences

Allelic variants are displayed as changes to publicly available reference sequences using identifiers that refer to a specific sequence accession and version. The PharmVar database stores variants on the following types of reference sequences:

Type	Example	Description
Locus sequence	NG_008376.3, LRG_303	RefSeqGene sequence or Locus Reference Genomic Sequence (LRG) if one exists for gene
GRCh37	NC_000022.10	Chromosomal reference from Genome Reference Consortium Human Build 37 (GRCh37)
GRCh38	NC_000022.11	Chromosomal reference from Genome Reference Consortium Human Build 38 (GRCh38)
mRNA sequence	NM_000106.5	RefSeq transcript sequence

In some cases, additional sequences may be listed for legacy reasons (e.g. M33388 for *CYP2D6*). These are determined on a case by case basis.

Locus sequences are genomic sequence including introns and exons and flanking regions of variable length, usually 5kb upstream and 2kb downstream. The Locus Reference Genomic (LRG) project (http://www.lrg-sequence.org) aims to designate a RefSeqGene (https://www.ncbi.nlm.nih.gov/refseq/rsg/) sequence as the stable, fixed sequence for each gene. PharmVar uses LRGs when they exist for a gene but will use the latest RefSeqGene sequence in all other cases. RefSeq transcript sequences reported will be the reference standard transcript appearing in the fixed annotation of the LRG or of the RefSeqGene sequence. Note that this transcript version may not be the latest version available from RefSeq.

Gene Identifiers

Each gene is listed using its official symbol and full name as specified by the HUGO Gene Nomenclature Committee (HGNC) along with common aliases and synonyms for the gene. Genes are listed with their identifiers in public databases such as NCBI and PharmGKB to facilitate cross-referencing.

Reference SNP ID numbers

rs ID numbers have been annotated in PharmVar using dbSNP build 152. As dbSNP is continually updating the database, PharmVar annotations may not reflect the most current dbSNP entries. PharmVar will periodically update rs annotations.

Coordinates

PharmVar provides variant coordinates using two different numbering conventions. **Sequence Start** numbering counts from the start of the sequence with the first nucleotide of the sequence being position 1 and incrementing sequentially to the end of the sequence. **ATG Start** numbering uses the A in the initiation start codon of the gene as position 1 and increments sequentially in the 3' direction. Positions before the ATG start are listed as negative numbers. In this scheme, there is no position 0; the base immediately preceding the ATG start is numbered -1. **ATG Start** numbering is provided for locus sequences only (i.e. not for chromosomal sequences). NOTE: this numbering is similar to the recommendations provided by the Human Genome Variation Society (HGVS) for coding sequences, however differs significantly by the inclusion of intronic sequences.

Substitution variants are listed with the position of the substituted nucleotide along with the reference base and substituted base. **Deletion variants** are listed with the start position as the first deleted base and the stop position being the last deleted base. **Insertion variants** are listed with the start position as the base immediately preceding the inserted sequence and the stop position listed as start position +1.

Insertion and deletion of nucleotides in a repeat or homopolymer sequence are listed using the **3' Rule** recommended by the HGVS. Using this rule, the inserted or deleted base will always be listed using the 3'-most position of the repeat sequence relative to the reference sequence.

		10	
reference	1	**GCACTAAAAAGCATGC**	**15**
variant		**GCACTAAAA-GCATGC**	

For example, given the above alignment of a variant sequence to the reference gene sequence, the deletion of one A is listed as the most 3' (right) A at position 10 rather than the most 5' (left) A at position 6.

For genes on the complement strand of the genome, an additional complication is introduced when listing the genomic position of the variant. If the gene sequence from the example above is displayed on the reference chromosome the reverse complement sequence will be displayed.

```
                                      110
reference       101     GCACTTTTTTGCATGC     115
variant                 GCACTTTTT-GCATGC
```

The original deletion at position 10 will be translated to position 106, however since this is now the first base of the repeat sequence, the deleted base will shifted to the right and reported as position 110 and listed with the deleted base being the complement ‘A’ instead of T.

The use of the 3’ most position may cause mismatches when comparing PharmVar variant positions to external resources such as dbSNP that prefer the 5’/leftmost convention. For example, the *CYP2C9*6* variant rs9332131 representing the deletion of a single A (AA/A-) is listed with a GRCh37 genomic position of chr10:96709040 in PharmVar, but is shown at position 9670939 at dbSNP. dbSNP is in the process of updating their notation to match the convention used by VCF files, i.e. where the position listed is the first position of the repeat sequence that is affected. This notation will not match the HGVS convention which is used by PharmVar.

Transcript Sequence Coordinates

PharmVar provides variant positions on RefSeq transcript sequences. In this view, intronic variants will not be displayed since they do not exist within the sequence. Additionally, note that some transcript sequences do not contain an annotated 5’-UTR (e.g. *CYP2C19*/NM_000769.1). In this case, **ATG Start** and **Sequence Start** positions will be identical.

Protein coordinates

PharmVar maps amino acid positions relative to the genomic RefSeq rather than to the putative amino acid sequence produced by a variant allele. If an allele carries a 3-nucleotide deletion causing the deletion of an amino acid, the amino acid positions following that deletion will still be reported in their original positions on the reference sequence without adjustment. For example, *CYP2D6*109* is characterized by a Lysine deletion at protein position 281; this variant also carries a Val to Met substitution which is reported as amino acid position 338 corresponding to the reference protein sequence and not 337 which would be the position in the variant allele protein sequence.

Frameshift variants

Variant positions are listed relative to RefSeq sequences. Variants downstream from an insertion or deletion variant will be listed with their positions on nucleotide reference sequences without adjustment even though the insertion or deletion will technically change the position of the downstream variant. For example, given the reference sequence reads TACG – a deletion of the T will be listed as position 1 while a substitution of G>C will be listed as position 4 even though the change would be at position 3 in the variant allele.

Amino acid changes after an indel are mapped to the genomic RefSeq and do not reflect any frameshifts caused by the indel. As an example, 3184G>A on *CYP2D6*109* is shown as V338M (and not V337M despite the L281del further upstream) (also see section on Protein coordinates above). Another example is *CYP2D6*4*. The 1847G>A SNP on causes a splice defect that shifts the reading frame. Impact for SNPs downstream of 1847G>A are shown independently of the frameshift, however. Therfore, 4181G>C is shown as S486T.

PharmVar ID

Haplotypes are identified by by PharmVar IDs (PVIDs), unique numeric identifiers similar to dbSNP's rs IDs. While star allele names are driven by functional grouping, they are not guaranteed to be permanent and may change over time as more information becomes available. For example, if an allele's star designation is updated to a new star number, the PVID of the haplotype remains constant and will not change. In contrast, if a haplotype definition changes (e.g. through the addition or removal of variants) a new PVID will be assigned. If the original PVID represents a haplotype definition that is still valid, it will remain in the database; otherwise, it will be retired.

For example, if the SNP at -899 of CYP2C19.004 (currently believed to have no functional consequence) were shown to cause a significant change in transcription levels leading to decreased function, this allele would be assigned a new star name. The PVID of this allele, PV00421, would stay the same indicating that there were no changes regarding the sequence variants defining this haplotype.

The PVID will change, however, if a new submission provides additional information that changes the haplotype definition itself. For example, the *CYP2D6*1.003* haplotype definition based on exon sequencing only contains only an exonic 1979C>T SNP (and has been rated as having a limited level of evidence). If a new CYP2D6 submission based on sequencing of the full locus were to describe a haplotype that carries 1979C>T and a novel intronic SNP, the haplotype definition of *CYP2D6*1.003* would be revised to include the intronic SNP. In this case, the allele name would not change, but the revised haplotype would receive a new PVID and the initial PVID will be retired.

PharmVar PVIDs are now searchable through the PVID lookup function (in menu bar). Through this function any PVID can be looked up and retired PVIDs tracked.

Core Allele Defintions

For many alleles there are a growing number of haplotypes, often referred to as suballeles, that share one or more 'key' defining sequence variant(s) (see ALLELE DESIGNATION AND EVIDENCE LEVEL CRITERIA document for details). While suballele information is valuable for e.g. the design and interpretation of test results (sequence or genotype-based alike), the distinction of suballeles is not necessary for phenotype prediction as **all alleles under a star number are**

believed to be functionally equal. Thus, even if a test is capable of distinguishing suballeles, these are generally simply reported as e.g. *CYP2D6*2* or **4,* etc.

PharmVar and the PharmGKB have collaboratively developed core allele definitions. Only sequence variations which change an amino acid or impact function by changing expression levels or interfere with splicing and are present in **all** suballeles within an allele group, are part of the core allele definition. **With this rule-based system suballeles are collapsed into a single 'core' definition representing all suballeles categorized under a star (*) number.**

Core alleles have their own unique PVID and can be tracked through the PVID lookup function. Note that all suballeles are identified by a numeric extension (e.g. .001, .002, etc.) core alleles do not have an extension. Thus, all PVIDs linked to a haplotype without an extension are core allele defintions.

Core allele defintions and the **Comparative Allele ViewEr (CAVE)** are currently available for *CYP2B6*, *CYP2C9, CYP2C19* and *CYP2D6*. More details and gene-specific examples are provided in each gene's GENE INFO document.