Published in final edited form as:

Clin Pharmacol Ther. 2015 February ; 97(2): 116–119. doi:10.1002/cpt.15.

Enabling the curation of your pharmacogenetic study

E.M. McDonagh¹, M. Whirl-Carrillo¹, R.B. Altman^{1,2}, and T.E Klein¹

¹Department of Genetics, Stanford University, Stanford, California, USA

²Department of Bioengineering, Stanford University, Stanford, California, USA

Abstract

As pharmacogenomics becomes integrated into clinical practice, curation of published studies becomes increasingly important. At the Pharmacogenomics Knowledgebase (PharmGKB; www.pharmgkb.org), pharmacogenetic associations reported in published articles are manually curated and evaluated¹. Standard terminologies are used, making findings uniform and unambiguous. Lack of information, clarity or standards in the original report can make it difficult or impossible to curate. We provide 10 rules to help authors ensure that their results are accurately captured and integrated.

Ten rule guide to enabling your publication to be curated

Numerous groups have raised the issue of standardization for reporting genetic variants and have established recommendations, guidelines and databases to aid in this effort. Among them are the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) which provides approved symbols for genes², and the Human Genome Variation Society (HGVS) which has established guidelines for sequence variation nomenclature^{3,4}. Many of the major pharmacogenes have nomenclature committees and/or databases that define which variants contribute to known alleles or haplotypes and provide mapping to reference sequences⁵.

The collection and evaluation of associations between genetic variants and drug responses is important, not only for determining clinically-actionable variants but for furthering research in the field of pharmacogenomics and contributing to future drug discovery/repurposing efforts. Standardization of genetic variant nomenclature, providing mapping information, making association analysis and results clear can aid curation of published results and thus access for the wider community.

PharmGKB (www.pharmgkb.org) is a comprehensive publicly-available resource that collects and curates knowledge about the impact of human genetic variation on drug response¹. We manually curate associations between genetic variants and drug responses

Correspondence to Dr Teri E. Klein, PhD, Department of Genetics, Stanford School of Medicine, Shriram Center, 443 Via Ortega, Room 213, Stanford, CA 94305-4125, USA. Tel: +1 650 736 0156; fax: +1 650 725 3863; teri.klein@stanford.edu.

Conflicts of interest

RBA is a founder, stockholder and scientific advisory board member to Personalis, Inc. TEK is a scientific consultant and stockholder in Personalis Inc.

reported in published articles and add them to the knowledgebase using numerous resources, including those mentioned above. Lack of clarity or absence of information is unfortunately common in the literature, making it difficult or impossible to curate a particular association. A publication should provide enough information for others to replicate the study, and allow for the results to be curated and compared against the findings of other studies. Attempts have been made to address such issues in certain areas and ensure publications provide sufficient information to enhance transparency, normalization of procedures and enable comparisons between studies, for example; the rules outlined by the “Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines⁶. Below, we provide a checklist of 10 rules that can be followed to better enable curation of your publication and make the findings accessible to a wider community (see Table 1).

1. Use standard gene nomenclature

HGNC-approved symbols² for human gene names should be used so that they cannot be mistaken for other proteins/genes. For example, “MDR1” may be used to refer to the *ABCB1* gene, however “MDR1” is also a synonym for the *TBC1D9* gene, thus using the HGNC approved symbol “*ABCB1*” avoids confusion.

2. Provide dbSNP rsIDs, a reference sequence and mapping information for genetic variants

Reference sequence identification (rsID) numbers are identifiers for a particular genetic variant established by the National Center for Biotechnology Information Short Genetic Variations database (dbSNP)⁷. Using an rsID enables a genetic variant to be referred to, is linked to information in dbSNP including genomic mapping in the form of HGVS³ nomenclature (described below), and allows readers to identify the same variant across multiple articles. RsIDs can be identified by several means (Table 2). If there is no existing rsID for a variant, submit the variant to dbSNP⁷ to obtain one.

It is also critical to provide a reference sequence for mapping the positions of genetic variants. We suggest using HGVS nomenclature³, as this encompasses the reference sequence, position of the polymorphism in that reference sequence and the allele change for a particular variant (*e.g.*, example provided in Table 1). NC_#, NT_#, NG_#, NM_#, and NP_# provided by dbSNP are NCBI reference sequences for chromosome, genomic scaffold, gene, mRNA and protein, respectively, and if referred to in a publication, can be found in the NCBI database using a simple search⁸. For more information on recommendations for the description of sequence variants, refer to the HGVS website^{3,4}, and the Mutalyzer website which contains tools that can help establish the correct nomenclature⁹.

3. Clearly state which allele/genotype of the variant is associated with the phenotype and the *direction* of the association

For example “allele A was associated with increased drug plasma concentrations compared to allele G”. This may seem obvious, but some studies simply report an association with a SNP *e.g.* “rsID# is associated with increased drug plasma concentrations” making it impossible to determine which allele or genotype is associated with the phenotype, or

“rsID# is associated with plasma concentrations” making it impossible to determine whether allele A is associated with increased or decreased plasma concentrations. If increased plasma concentrations were associated with the minor allele, specify which allele this was (rule 5).

4. Clearly state population size, ethnicity, gender and drug regimen

Sometimes genotyping is not completed for all variants in all individuals in the cohort, or patient subsets are not included in particular analyses. This information should be clearly stated for each association. Population characteristics such as ethnicity are also important to include as lack of association may be because the allele frequency within this population is too small for the variant to have a detectable effect in that study size.

5. State the minor allele for the variant in the given population/cohort

Authors often refer to the minor allele without explicitly stating what that allele is. However, the minor (less frequent) allele in one population may be the major allele (more frequent) in a different population. An example is rs1858923 in the *ABCB1* gene (NC_000007.14:g.87591900A>G). The A allele is found at a frequency of around 90% in Luhya individuals (Webuye, Kenya); however, the G allele is found at a higher frequency than the A allele in Gujarati Indians (Houston, Texas), and in several other diverse populations each allele is close to 50%¹⁰. In this case, stating “the minor allele” is not sufficient to clearly indicate which allele is being referenced. Instead, both the allele and the population should be clearly stated, as in “the minor allele, A, in Gujarati Indians.”

6. Clearly state on which chromosomal strand the alleles are reported

For consistency, PharmGKB reports all associations with alleles on the positive chromosomal strand; alleles for variants in genes on the minus chromosome (as indicated by NCBI Gene⁸) are complemented (the NCBI NC_# reference sequence shows alleles on the positive strand⁷). It can become difficult to ensure the correct allele for a particular association for A/T or G/C SNPs, especially if codon information is not applicable, for example rs12201199 (NM_000367.2:c.419+94T>A) within an intron of the *TPMT* gene. A statement such as “the A allele is associated with increased drug response” in a gene on the minus chromosomal strand may be referring to the A allele in the gene (and therefore the T allele on the positive chromosomal strand), or the A allele on the positive chromosomal strand. It is therefore not possible to have confidence in which allele is associated with the given phenotype, whereas “the A allele (positive chromosomal strand) is associated...” is clear.

7. State all genetic variants that were screened (including SNPs, indels, copy number variations and structural variations)

It can be important to know whether certain variants were screened for when stating that ‘wildtype’ individuals were considered those without variant alleles of a group of polymorphisms (see rule 8). These individuals may not necessarily be ‘wildtype’ if only a few variants were examined. This is particularly important in the setting of sequencing versus genotyping, where the former may interrogate many more potential variations than the latter. In addition, it should be stated whether other types of variants were evaluated (and

therefore could have been discovered) such as copy number variations, structural variations, and indels. The tissue source of the DNA should also be specified. Large amounts of information, such as sequencing or SNP platform details, could be included in supplemental material.

8. Specify how * alleles or haplotypes were defined

Even when referencing standard nomenclature (for example “* alleles were determined as defined by the *NAT* gene nomenclature committee”⁵), it is good practice to provide full details in the current publication (see Table 1). For alleles consisting of more than one variant, it is important to describe what methods were used to determine that these were on the same chromosome (cis) rather than one variant on each chromosome (trans). For example, sequencing could enable the identification of rs1801280 c.341 allele C and rs1799929 c.481 allele T on the same allele to determine *NAT2**5A.

9. Report which genotypes/diplotypes were found in the study population and map them to phenotype groups

This reporting makes clear which genotypes/diplotypes were actually present in the study population (for example refer to Table 1; the *NAT2**7A allele was not found).

10. State the statistical tests used for each association analysis. Include methods used for multiple hypothesis correction

When curating an association, statistical information is critically important to assess the strength of the association. Often it is difficult to determine which statistical test was carried out for each reported association and why certain tests were chosen. When curating an association, a correction term could be applied for multiple testing of genetic variants and parameters; however, it is often not clear how many genetic variants were originally screened for and how many different tests were applied, or, whether multiple testing has been accounted for already.

Conclusion

There are multiple benefits for all by following these 10 simple guidelines. By enabling published associations to be curated and integrated, they become part of the larger collection of curated PGx literature. Results become accessible to a broad audience, including students, bioinformaticians, researchers and clinicians in the context of the published works. This contributes to further research studies, such as genome interpretation, text mining, and drug discovery. Ultimately, this curated information may contribute to the clinical interpretation of pharmacogenetic associations, and may accelerate the implementation of pharmacogenetic-based drug dosing¹.

Acknowledgements

This work is supported by the NIH/NIGMS (R24 GM61374; TE Klein and RB Altman, PIs). We would like to acknowledge the work of others in this area, particularly the Pharmacogenomics Research Network (PGRN) Translational Pharmacogenetics Project, CDC Nomenclature Group and the ClinGen PGx Working Group.

References

1. Whirl-Carrillo M, et al. Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics*. 2012; 92:414–417. [PubMed: 22992668]
2. Gray KA, et al. [Genenames.org](http://www.genenames.org): the HGNC resources in 2013. *Nucleic acids research*. 2013; 41:D545–D552. [PubMed: 23161694]
3. [(accessed June 10th 2014)] Human Genome Variation Society (HGVS). <http://www.hgvs.org/>
4. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human mutation*. 2000; 15:7–12. [PubMed: 10612815]
5. [(accessed 10th July 2014)] Links to human pharmacogene variant nomenclature committees or resources for genetic variant names can be found here. <https://www.pharmgkb.org/search/browse.action?browseKey=haplotypeGenes>
6. Bustin SA, et al. The need for transparency and good practices in the qPCR literature. *Nature methods*. 2013; 10:1063–1067. [PubMed: 24173381]
7. [accessed June 10th 2014] NCBI Short Genetic Variations (SNV) database, also known as dbSNP. <http://www.ncbi.nlm.nih.gov/projects/SNP/>
8. National Center for Biotechnology Information (NCBI). 8600 Rockville Pike, Bethesda MD, 20894 USA: U.S. National Library of Medicine; URL: <http://www.ncbi.nlm.nih.gov/> [accessed June 10th 2014]
9. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Human mutation*. 2008; 29:6–13. [PubMed: 18000842]
10. International HapMap, C. The International HapMap Project. *Nature*. 2003; 426:789–796. [PubMed: 14685227]

Table 1

Checklist of ten rules to help enable your publication to be curated.

Check	Rule #		Example
	1.	Use HGNC approved symbols for gene names.	<i>ABCB1</i> (rather than <i>MDR1</i>).
	2.	For all genetic variants, provide a dbSNP rsID, and a reference sequence with mapping information.	For example, rs1801280 in the <i>NAT2</i> gene can be described as NM_000015.2:c.341T>C (the nucleotide reference sequence: position of the polymorphism within this cDNA sequence and it is a T to C allele change).
	3.	Clearly state which allele/genotype of the variant is associated with the phenotype and the <i>direction</i> of the association.	Allele A was associated with increased drug plasma concentrations compared to allele G.
	4.	Clearly state population size, ethnicity, gender and drug regimen.	All patients were female, of Han Chinese ethnicity; n=10 were treated with allopurinol and n=5 treated with carbamazepine.
	5.	If referring to the minor allele of a variant, state which allele this in and for which given population/cohort.	The minor allele, A, in Gujarati Indians, was associated with increased drug plasma concentrations compared to allele G.
	6.	Clearly state on which chromosomal strand the alleles are reported.	The A allele (positive chromosomal strand) is associated with increased drug plasma concentrations compared to allele T.
	7.	State all genetic variants that were screened (including SNPs, indels, copy number variations and structural variations), and specify the tissue source of DNA. <i>This information could be provided in the supplement.</i>	rs1799929 NM_000015.2:c.481C>T, rs1799930 NM_000015.2:c.590G>A, rs1799931 NM_000015.2:c.857G>A, rs1801280 NM_000015.2:c.341T>C. DNA was isolated from PBMCs purified from whole blood.
	8.	Specify how * alleles or haplotypes were defined.	<i>NAT2</i> *5A was defined as rs1801280 c.341 allele C and rs1799929 c.481 allele T, <i>NAT2</i> *6B was defined as rs1799930 c.590 allele A, <i>NAT2</i> *7A was defined as rs1799931, c.857 allele A. The <i>NAT2</i> *4 “wildtype” allele was defined as <i>NAT2</i> rs1801280 c.341 allele T, rs1799929 c.481 allele C, rs1799930 c.590 allele G, rs1799931 c.857 allele G (no variant alleles at the positions genotyped).
	9.	Report which genotypes/diplotypes were found in the study population and map them to phenotype groups.	Slow acetylators were defined as two ‘slow’ alleles of any combination of <i>NAT2</i> *5A, <i>NAT2</i> *6B, <i>NAT2</i> *7A alleles. Rapid acetylators were defined as one or two <i>NAT2</i> *4 alleles. Five individuals were grouped as slow acetylators (genotypes present: n=2 <i>NAT2</i> *5A/ <i>NAT2</i> *5A, n=1 <i>NAT2</i> *5A/ <i>NAT2</i> *6B, n=2 <i>NAT2</i> *6B/ <i>NAT2</i> *6B), 10 individuals were grouped as rapid acetylators (genotypes present: n=5 <i>NAT2</i> *4/ <i>NAT2</i> *4, n=3 <i>NAT2</i> *4/ <i>NAT2</i> *5A, n=2 <i>NAT2</i> *4/ <i>NAT2</i> *6B).
	10.	State the statistical tests used for each association analysis. Include methods used for multiple hypothesis correction.	Mann Whitney U Test. All p-values provided were corrected for multiple comparisons by Bonferroni correction.

Table 2

Tips on how to identify a rsID for a genetic variant.

1.	Search dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) for the gene name. Select 'Limits' for the search to narrow down the search (e.g. SNP class). Select 'Display Settings' and display 200 items per page. Search for mapping, position or amino acid change information in the list of variants. Use Gene View to zoom in if position information is available to identify possible rsIDs.
2.	Search PharmGKB (www.pharmgkb.org) for the gene and view the 'PGx Research' tab for variant information. Search for mapping, position or amino acid change information.
3.	Search UCSD genome browser (http://genome.ucsc.edu).
4.	Search for the gene on SNPperCHIP (http://snpper.chip.org/).
5.	Search for the gene on SNPedia (http://snpedia.com/index.php/SNPedia).