The Star-Allele Nomenclature: Retooling for Translational Genomics

JD Robarge[1], L Li[1], Z Desta[2], A Nguyen[2] and DA Flockhart[2]

The star-allele nomenclature is the result of efforts to standardize genetic polymorphism annotation for the cytochrome P450 genes. As clinical pharmacogenetic testing becomes widespread, it is important that this system effectively communicate a patient's genotype and predicted clinical phenotype. As genomics research expands, it is equally important that the system remain a valuable tool for the wider community of genetic researchers to exploit our ever-improving ability to catalog variability in the human genome.

Phenotype-to-genotype approach

Uncovering the role that polymorphic cytochrome P450 isoforms play in variable drug response began with observations of unusual, inherited drug responses. Attempts to relate this variability to hereditability began with the observation of defective oxidation of drugs by CYP2D6. This polymorphism was independently discovered in three laboratories (**Box 1**, refs.1–4).

The discovery of these genetic polymorphisms was thus made possible by astute observations of unusual drug response by alert clinicians. It was driven by observations of phenotype for which the underlying molecular genetic causes were subsequently elucidated—a phenotype-to-genotype approach that remains valuable. An expanded list of probe drugs for cytochrome P450 enzyme activity is available, and these probes are widely used by researchers and in drug development to identify the poor-metabolizer, extensive-metabolizer, ultrarapid-metabolizer, and even intermediate-metabolizer phenotypes of these enzymes. Although this phenotypic approach does not capture the molecular variants involved, it remains a valuable research tool and a useful means of quantitating gene–environment interactions for these enzymes.

The star-allele nomenclature

As the genetic basis for variable drug metabolism became increasing clear, concurrent with improvements in genotyping and sequencing technologies, genetic variants contributing to poor-metabolizer phenotypes were discovered at a rapid pace using probe drugs. Forward thinkers in the field, led by Dan Nebert and Magnus Ingelman-Sundberg, recognized the need for careful curation and annotation during this explosion in genetic data.[5] For the cytochrome P450 enzymes and a number of other metabolic gene families, nomenclature committees took the role of curators of existing and newly described variants, standardizing their description and allelic designation. The Cytochrome P450 (*CYP*) Allele Nomenclature Committee maintains allele nomenclature for genetic polymorphisms in 25 of 58 known human *CYP* genes.[6] Other websites maintain similar pages for pharmacologically important genes, such as UDP-glucuronosyltransferase (*UGT*) (http://galien.pha.ulaval.ca/alleles/alleles.html) and *N*-acetyltransferase (*NAT*) (http://louisville.edu/medschool/pharmacology/NAT.html), providing a common allelic nomenclature widely referenced in practice and publication. These websites list gene alleles with observed nucleotide changes, a single-word classification of functional effect, and links to publications detailing their discovery or functional characterization.

Definition, naming, and inclusion criteria. In star-allele nomenclature, *1 is the designated reference sequence with which polymorphic sites are compared. This is usually the first sequence described that codes a functional protein product, and may not be the most common allele in every ethnic population. Subsequently, a unique number (*e.g.*, *CYP3A5*3*) is assigned when a novel variant is identified with a nucleotide change that leads to an amino acid substitution or is shown to affect transcription, splicing, translation, or post-transcriptional or post-translational modification. Nonfunctional nucleotide changes thought to exist on the same chromosome, or to be inherited with a named star allele, are not given new numbers but are defined by additional letters (*e.g.*, *CYP3A5*3B, CYP3A5*3C*), where the principal star allele is designated by an A (*e.g.*, *CYP3A5*3A*). If multiple variant alleles existing on the same chromosome are shown to have a functional effect on the protein, in a context where no single polymorphism has an effect, a new allele number is designated (*e.g.*, *CYP2D6*17*).

Since 1999, the Cytochrome P450 (*CYP*) Allele Nomenclature Committee has maintained *P450* allelic designations following a peer-review process. The committee has exceeded its initial goal of a standard naming scheme for genetic variants discovered in the *CYP* genes. Its function-centric focus has been an important component in the advancement of human genetics in pharmacology, allowing studies of phenotype–genotype correlation to

[1]Division of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, USA; [2]Division of Clinical Pharmacology, Department of Medicine, Indiana University School of Medicine, Indianapolis, Indiana, USA. Correspondence: JD Robarge (jrobarge@iupui.edu)

doi:10.1038/sj.clpt.2007.6100284

be translated between researchers. A standardized nomenclature has been clearly embraced by the pharmacology community, which can be seen through its widespread adoption in publication since its introduction. However, because this system was implemented before the rapid expansion of genomic science, it has inherent limitations. There is a pressing need to make this system consistent with the nomenclatures common across the wider field of genetics by incorporating recent genomics data and annotation and by further re-evaluating and expanding the definition and description of a functional nucleotide change.

Box 1 Discovery of the CYP2D6 polymorphism.

- In 1967, Sjöqvist and co-workers in Sweden identified two groups of patients based on the plasma concentrations of nortriptyline and desimipramine after a fixed dose, and later showed that this difference was genetic in nature.[1]
- In the mid-1970s, Smith's group in the United Kingdom discovered that 4-hydroxylation of debrisoquine in humans is polymorphic, and that the poor metabolizer phenotype is inherited as a single autosomal recessive trait.[2] The terms "extensive metabolizer" and "poor metabolizer" described at that time continue to be used today.
- In 1975, Eichelbaum _et al._ identified the sparteine polymorphism through the observation of two patients with defective metabolism.[3]
- Ten to fifteen years after these differences in phenotypes were discovered, the genetic basis underlying the bimodal debrisoquine–sparteine phenotype was elucidated by collaborative work between Urs Meyer and Frank Gonzalez[4] in which they cloned the human CYP2D6 cDNA, documented altered RFLP patterns in slow metabolizers of debrisoquin, and later identified what we now refer to as the _CYP2D6*3_ and _CYP2D6*4_ alleles (star (*) alleles are defined in the text).

Annotation of genetic polymorphisms. Originators of the _P450_ allelic designations could not have anticipated the post-genome explosion of genetic polymorphism data. Advances in sequencing and genotyping technology have made high-throughput discovery and characterization of genetic variants in humans feasible and routine. Currently, there are 11,811,594 reference single-nucleotide polymorphisms (RefSNPs) described in dbSNP (build 127), with 5,689,286 validated, and 5,028,168 mapping to genes (http://www.ncbi.nlm.nih.gov/SNP). Recent studies of copy-number polymorphisms and structural variation have revealed a large amount of genetic polymorphism not accounted for by commonly assayed SNPs or by variable number of tandem repeat polymorphisms. Whole-gene resequencing approaches have generated notably more comprehensive data on genetic variation of some genes encoding drug-metabolizing enzymes.

Although the use of the star nomenclature is widespread, it is limited in its ability to present and annotate genetic variation. A focus on functional genetic variation in exons and potential splice variants has led to an artificial divide between the star nomenclature and high-density polymorphism data being generated elsewhere. In the star-allele tables, nucleotide variant positions and their respective amino acid substitutions are annotated relative to GenBank and SwissProt reference sequences when there is a functional association. Links to dbSNP are provided for functional SNPs with refSNP ids. Conversely, polymorphisms without a clear or known functional role (_e.g._, that do not lead to a non-synonymous amino acid change) are labeled relative to only the GenBank reference sequence. Apart from those variants with links to dbSNP, no variant can be easily mapped to the NCBI human genome assemblies, making it difficult to integrate the star-nomenclature variants with information from widely used tools or databases such as the University of California–Santa Cruz Genome Browser (http://genome.ucsc.edu).

Annotation of population distribution. While researchers familiar with the star nomenclature often know the ethnic distribution of star alleles for a given gene, star-nomenclature tables make the explicit assumption that all variants exist in any population and make no indication as to their relative frequencies. Conveying the ethnic-specific distribution of all gene polymorphisms as well as named star alleles would aid researchers and clinicians in use and interpretation of the star-allele tables.

Annotation of phenotypic effects. Pharmacogenetic phenotypes are as broad as the organisms and systems in which they are measured and the tools used in measurement. A measured phenotype may be dependent on temporal and environmental effects, as well as on genetic effects in drug-metabolizing enzymes or drug targets. Observing a genotype effect in this multidimensional phenotypic space is already a difficult research task; extrapolating to potential clinical utility is far more challenging. Use of genetic markers as biomarkers for clinical outcomes should be considered only after appropriate clinical validation, but outside of such trials, precise descriptions of genotype and related clinical associations need to be clearly described for any "functional" genetic variant. Genetic variants that compose the star alleles are unique in population genetics due to their history of discovery. In genetic association studies, genetic variation serves as a marker to determine the effect of genotype on a phenotype. Conversely, star alleles were discovered based on their functional effects. Importantly, any organization of the star alleles must also be accompanied by the phenotypic descriptions by which they were discovered. Although the star system as represented on the Web provides limited annotation of phenotypic effects, often genetic effects measured _in vitro_ and _in vivo_ are described in non-quantitative terms: increased, decreased, severely decreased, abolished, none, normal, negligible, or unknown function. As a supplement to functional variant annotation, an inventory of clinical studies used, including sample and effect sizes, would allow the research community to fully leverage this resource.

Recently, through independent curation efforts by the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB), links between star nomenclature and current genomic annotation are being made in the "VIP Annotated PGx Gene" pages,[7] where highly annotated polymorphisms are presented with population frequencies, textual description of effects on molecular or clinical phenotype, PubMed IDs of relevant publications, and other information. We hope that the star-nomenclature and PharmGKB sites can reach a consensus on the presentation of this information while preserving the strengths of each site.

Star alleles and haplotypes. Descriptions of patterns of linkage disequilibrium between genetic markers and haplotype structure have appeared concurrently with gene and genome-wide polymorphism discovery efforts. A haplotype, or "haploid genotype," is the genetic constitution of an individual chromosome. A haplotype can refer to a gene-centric locus or an entire chromosome. In diploid organisms, a "haplotype pair" refers to the specific set of alleles "observed" on both chromosomes for a given number of polymorphic loci. However, haplotypes must be molecularly determined or inferred using computational methods. While current molecular haplotyping methods are cost restrictive and laborious, genotyping followed by computational haplotype inference is widely used as an efficient method for haplotyping. Computationally inferred haplotype pairs are statistically associated alleles that have been phased (assigned to both chromosomes) on the basis of observed genotype data.

For a lone compound-heterozygous individual, the underlying haplotype pair cannot be strictly determined without molecular haplotyping. Alternatively, genotype data from family members—or a sample of individuals thought to have a similar ethnicity and thus to share a common genetic ancestry—can help determine probabilities of various haplotype pairings given an individual's multi-locus genotype. With any compound-heterozygous individual, there is phase ambiguity and potential error associated with computationally determining haplotypes. Computational inference of haplotype phase in family-based studies utilizes a known pedigree structure to trace the transmission of chromosomal regions from parents to offspring.[8] In pharmacology, obtaining phenotypic data from family members is not often possible and thus DNA is rarely collected from an individual's parents or offspring. In studies of unrelated individuals, haplotypes can be efficiently estimated from unphased genotypes using expectation-maximization or Bayesian-based computational methods.[9] These algorithms estimate population haplotype frequencies and can probabilistically assign haplotype pairs to individuals based on observed and unphased multi-locus genotypes in a population sample. Computational approaches for haplotype determination, which are highly prevalent in population genetics research, are a mathematically optimal method for describing allelic phase of multiple genetic polymorphisms. Computational approaches have recently been used to describe the ethnic-specific distribution of haplotypes in genes encoding important drug-metabolizing enzymes, such as the CYP3A cluster. Using a similar approach, VKORC1 haplotypes and CYP2C9 star alleles have been integrated to predict warfarin international normalized ratio maintenance in a clinical setting.[10]

In its current representation, a discrete star allele represents either a single genetic variant or a haplotype. Some studies have used molecular methods such as allele-specific polymerase chain reaction (PCR), limited-dilution PCR, denaturing high-performance liquid chromatography, cloning, or computational approaches to characterize haplotypes of varying length (*i.e.*, that encompass a range of neighboring polymorphisms) and to determine new alleles. Star alleles have often been determined through a historically iterative approach, whereby new alleles are defined for normal individuals or for individuals with phenotypes not predicted by their previously assigned star alleles (star genotype) using genotyping or re-sequencing data. Further complexity in star alleles for a given gene stems from the fact that not all studies have genotyped or sequenced the same polymorphisms. The accuracy in assignment of a new star allele (equivalent to determining the degree of allelic association between the new variant and all other known variants) is dependent on which variant alleles are interrogated, their frequencies in the populations studied, and the number of chromosomes sampled. Although extensive standards exist for naming new genetic variants in the *P450* nomenclature, there is no clear standard for communicating which alleles exist within which haplotypes, or how rigorous the methods used to do so should be (*e.g.*, how many samples of which ethnicity analyzed by what methods).

Two important and related issues need to be addressed if there is to be a seamless integration between star-allele and haplotype approaches and clinical use of either. First, it is not clear that the history of genetic polymorphism discovery, represented in the current *P450* nomenclature tables, has allowed accurate determination of the co-occurrence of individual genetic variants within the star alleles. For well-studied genes, such as *CYP2D6*, many named star alleles clearly exist in populations at frequencies that have been established. However, other haplotypes may exist that have not been described because of the complex history behind how each allele was determined. Second, in ongoing pharmacogenetic studies, star-allele tables are utilized to assign genetic variants that are observed to individual alleles, annotated as star-allele designations (*e.g.*, *4/*6). The assignment is both deterministic, meaning that it is established without uncertainty, and independent of ethnicity. Can we accurately determine an individual's haplotype pair (star-allele designations) observationally, without computational or molecular methods? This approach can be considered valid only under the assumption that alleles of the genetic markers tested for act as universal haplotype tags. The tagging ability of these variant alleles can be validated only for populations that have been investigated and not for unstudied or remote populations. Co-segregation of well-studied variant alleles in haplotypes may be sufficiently described in commonly studied populations. However, given the difficulty of estimating haplotype phase using even the best computational methods, star alleles assigned without error using star-allele tables should be viewed with caution.

A compromise between deterministic and probabilistic haplotype assignment needs future discussion. The consequence of incorrect assignment of haplotype alleles will simply be inaccurate predictions of clinical outcome. These may be of little moment, but they may also be very important in situations where key genes, important applications, or diseases are involved. Routine clinical use of genetic diagnostic tests is clearly approaching. We should therefore utilize methods in the clinic that most accurately determine an individual's genetic makeup and that can be used to clearly communicate a clinical consequence.

Recommendations

If pharmacogenetic biomarkers are to be of clinical value, genetic nomenclature of pharmacologically important genes must accurately represent underlying genetic variation and should clearly describe the association between this variation and important functional consequences. These might include *in vitro* assays, pharmacokinetic changes, or clinical outcomes as are used by the PharmGKB database. The phenotypic organization of the PharmGKB database could serve as a model for how phenotypes can be grouped and described.

We recommend that the following changes be made to the system and to its description on the star-alleles website.

1a. Annotating genetic polymorphisms. To enable full integration of the current star-allele nomenclature with ongoing genome-scale variant discovery, an effort needs to be made toward annotation of genetic variants with refSNP IDs and human genome build positions. This will allow variants to be easily cross-referenced to a wealth of available human genetics data.

1b. Determining gene alleles. Resequencing studies of drug-metabolizing enzymes have confirmed previously named gene alleles but have also discovered unknown haplotypes. The number of gene haplotypes increases as the number of known genetic variants in a gene rises. To increase transparency and ensure accurate and valid phasing of gene alleles, genotype and sequencing data used for star-allele determination should be clearly described and made accessible to the broader research community. Web-based availability of primary genotyping data will allow haplotype determination to be validated and new methods of determining allelic phase to be tested and compared.

2. Cataloging cytochrome P450 substrates. The system should be linked at key points to an easily accessible source of updated and peer-reviewed information containing data on drugs that are substrates for specific enzymes, such as the site that we maintain at http://www.drug-interactions.com. Such a capability would greatly improve the clinical accessibility and utility of the system, while simultaneously allowing an ongoing exchange of data on metabolic elimination. Ideally such a system would also enable an assessment through the literature of the proportion of elimination documented to occur via any given enzyme.

3a. Determination of functional variant effects. The star-nomenclature tables should be accompanied by focused descriptions of research findings that support the association between specific variants and functional consequences for each gene. The currently provided information is inadequate for this purpose and would benefit from the inclusion of sample descriptions (*e.g.*, ethnicity and sample size) and molecular and computational methods used in determining associations. Moreover, it should clearly link to the science underlying associations with reported therapeutic outcomes.

Where possible and supported by evidence, a scoring system based on the expected phenotypic change given a genotype should be included, as suggested recently for CYP2D6. Statistical guidelines that can assign phenotype to a specific genotype are needed. The best scoring system should possess the highest functional predictive power based on genotype, and the chance of making false-positive statements should be kept at a pre-specified level.

3b. Communication of functional variant effects. Since clinical laboratories and clinicians increasingly use the star-allele system and website to decide on therapeutic strategies, we recommend that alleles be arranged on the website according to phenotypic activity. It would thus be possible for clinicians to easily access all alleles of low, normal, or high activity together with the evidence base that supports the classification. Although data derived from expression of variants *in vitro* and from human liver microsomes are of value, we believe the site should also contain evidence of *in vivo* replication when available, since many *in vitro* changes do not translate to clinically important effects.

Successful implementation of recommendations 1a, 1b, and 3a will depend on our ability to effectively combine results of genotyping, *in vitro*, and *in vivo* measurements from independent studies. An alternative approach may be taken to achieve these goals. The value of large-scale studies in which multiple genotypic or phenotype measures are obtained from a fixed biological sample is evident from the success of the HapMap (http://www.hapmap.org) and NCI-60 (http://discover.nci.nih.gov) projects. The strength of this approach is the ability to seamlessly integrate new and updated measures of phenotype and genotype with previous results. As opposed to these genome-wide approaches, pharmacogenetics may benefit from a similar but targeted scientific approach, in which genetic studies would focus on drug-metabolizing enzymes and substantial weight would be given to measured PK phenotypes.

Accumulating *in vitro* metabolic data along with genotyping and sequencing data using a large, ethnically diverse liver bank would be an optimal approach to the verification of new star alleles. Existing drug-metabolism data would serve to determine the effects of newly discovered genetic variants, while genetic effects on newly measured compounds could be queried with existing genotype data. A new resource of this depth would allow for precise characterization of genetic variants with small effects on enzyme function, and enable the combined effects of variants in multiple genes to be tested. This endeavor would require global logistical and financial cooperation among institutions, but it would be a powerful resource with which to address important questions that cannot be fully addressed in current *in vitro* pharmacogenetic studies of single drug–gene combinations.

Conclusion

The star nomenclature is a curated and widely used reference, and the *de facto* resource for genetic variation in *P450*s. To maximize the information content of this resource for future research, the star-allele website should include descriptions of research findings that support the association between specific variants and functional consequences for each gene. To further the clinical applicability of star-allele designations, allelic effects on functional consequences and therapeutic outcomes must be thoroughly described. We recognize that implementation of these proposals requires significant organizational work, as well as the support of the research and funding communities, but we believe that now is the time to begin this discussion. Our hope is that the genetic data underlying therapeutic decisions can ultimately be made as accessible and transparent as possible, so that the science underlying critical clinical decisions can be robustly and efficiently queried.

CONFLICT OF INTEREST

The author declared no conflict of interest.

1. Alexanderson, B., Evans, D.A. & Sjöqvist, F. Steady-state plasma levels of nortriptyline in twins: influence of genetic factors and drug therapy. *Br. Med. J.* **4**, 764–768 (1969).
2. Mahgoub, A., Idle, J.R., Dring, L.G., Lancaster, R. & Smith, R.L. Polymorphic hydroxylation of debrisoquine in man. *Lancet* **2,** 584–586 (1977).
3. Eichelbaum, M., Spannbrucker, N. & Dengler, H.J. Proceedings: N-oxidation of sparteine in man and its interindividual differences. *Naunyn Schmiedebergs Arch. Pharmacol.* **287**, Suppl:R94 (1975).
4. Gonzalez, F.J., Skoda, R.C., Kimura, S., Umeno, M., Zanger, U.M., Nebert, D.W. *et al.* Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* **331**, 442–446 (1988).
5. Nebert, D.W. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. *Pharmacogenetics* **10**, 279–290 (2000).
6. Sim, S.C. & Ingelman-Sundberg, M. The human cytochrome P450 Allele Nomenclature Committee Web site: submission criteria, procedures, and objectives. *Methods Mol. Biol.* **320**, 183–191 (2006).
7. Klein, T.E. *et al*. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* **1**, 167–170 (2001).
8. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet*. **58**, 1347–1363 (1996).
9. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet*. **68**, 978–989 (2001).
10. Sconce, E., Khan, T.I., Wynne, H.A., Avery, P., Monkhouse, L., King B.P. *et al*. The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood* **106**, 2329–2333 (2005).

See ARTICLE page 294

Protease Inhibitors as Immunomodulatory Drugs for HIV Infection

SG Deeks[1]

More than 20 drugs from four therapeutic drug classes are widely available for the management of human immunodeficiency virus (HIV) infection, with promising drugs from two new drug classes expected to be approved by the US Food and Drug Administration (FDA) in mid-to-late 2007 (Table 1). When used in combination, these drugs can lead to durable and perhaps indefinite suppression of viral replication.

Once the virus is controlled, peripheral $CD4^+$ T-cell counts often (but not always) increase, eventually reaching near normal levels after several years of therapy. In regions where "highly active antiretroviral therapy" (HAART) is available, HIV-associated opportunistic infections have become increasingly rare. The expected life span for patients with HIV infection who have access to and adhere to HAART is now measured in decades.

Despite clear success with these drugs, there remain limitations. Many patients have received sequential suboptimal regimens and harbor highly resistant HIV, or were infected with drug-resistant variants. Also, many of the most widely used antiretroviral drugs cause significant long-term toxicity (*e.g.*, lipid abnormalities, insulin resistance, abnormal fat distribution, and peripheral neuropathy). Fortunately, many of these problems are expected to be at least partially ameliorated by the recent development of more effective and safer drugs.

Perhaps the most significant unaddressed limitation associated with long-term antiretroviral therapy is the inability of these drugs to fully restore a normal immune system, at least in some individuals. Many patients receiving standard HAART fail to achieve normal

[1]University of California, San Francisco and San Francisco General Hospital, San Francisco, California, USA. Correspondence: SG Deeks (sdeeks@php.ucsf.edu)

doi:10.1038/sj.clpt.2007.6100205