Annotation of Genetic Polymorphisms in Drug-Metabolizing Enzymes (poster 1266/T)

Chia-Chien Chiang[1], Zhiping Gu[1], Shuang Cai[1], Rosane Charlab[1], Katherine Lazaruk[2], Alexander Levitsky[1], Chun-Hua Wan[1], Timothy Harkins[2], and Daniel Ingber[1]. 1) Applied Biosystems, 45 West Gude Drive, Rockville, MD 20850; 2) Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404

ABSTRACT

Many polymorphisms within the genes of drug-metabolizing enzymes (DME) have been shown to alter drug responses in individuals. A DME genotyping platform would allow researchers to screen for these polymorphisms and may serve as an aid to determine individualizing treatment choice. However, a comprehensive catalog of DME polymorphisms is lacking, presumably due to the challenges in facing inconsistent nomenclatures and the high-degree homology of DME-coding genes. Here we report a bioinformatics process to annotate DME polymorphisms based on the data collected from both public and proprietary databases. Our process involves mapping the flanking sequences of polymorphic sites to genome assemblies, clustering DME polymorphisms, and assigning functional classification to the polymorphisms based on their relative positions to DME proteins. Ambiguous results were inspected and resolved manually by experts. To help researchers identify the polymorphisms as reported by several locus-specific allele nomenclature committees, we assigned the allele nomenclature recommended by these committees using a pipeline that extracts and calculates the coordinates of DME polymorphisms on the reference sequences designated by the committees. The nomenclature is assigned to a variant when its locations and bases both match what were reported by the committees. We have so far annotated more than 3,000 protein-coding SNPs for 219 DME-related genes.

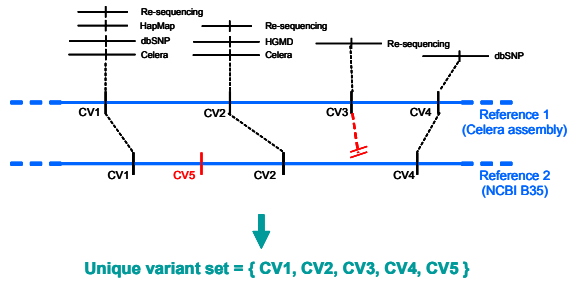
CREATION OF A NON-REDUNDANT GENETIC VARIANT SET

Table 1. Sources of genetic polymorphisms

Data Sources	Description
dbSNP	Public repository
HapMap	Public repository
HGMD	Public repository with special agreement
HGVbase	Public repository
Celera discovery pipeline	Comparison of aligned shot-gun and BAC-clone sequencing reads
Applera re-sequencing project	Comparison of aligned sequencing reads in the targeted regions

A non-redundant genetic variant collection was created using the data described in Table 1. The flanking sequences of the variants from public repositories were mapped to the reference assembly (Celera assembly R27) by the AB mapping software (ref.1). As illustrated in Fig. 1, the genetic variants of the above-mentioned sources were then clustered by a collocation process, which assigns the variants having the same start and end coordinates on the reference genome into the same cluster. Each unique cluster represents one non-redundant polymorphic site and was assigned with a unique Celera accession number (CV#). The CVs were tracked to NCBI Build 35 assembly by mapping. Due to the difference between the two reference assemblies, some variants may be located on only one of the two assemblies. For example (Fig. 1), CV3 found on Celera assembly does not map to Build 35. *Vice versa*, CV5 is instantiated based on the reported position on B35 of the dbSNP rs variant that does not map to Celera assembly.

Figure 1. Example to illustrate the process that creates non-redundant genetic variant set



SELECTION OF POLYMORPHISMS IN DME-CODING GENES

Approx. 220 genes were selected due to the various roles that they play in drug metabolism. Refer to 'Drug Metabolism Genotyping Assay Index' on the myScience web site (ref.2) for the complete list of DME genes that were selected. The CVs were selected if their coordinates on the reference assemblies are within the spans of the 5' untranslated or protein-coding regions of RefSeq transcripts (NM prefix) corresponding to the selected DME genes.

Table 2. Classification of selected variants

Variant Functional Type	Number	% of total
Protein-coding	3214	72.22
Mis-sense	2076	46.65
Nonsense	156	3.51
Frameshift	157	3.53
Splice-site	45	1.01
5' UTR	309	6.94
3' UTR	282	6.34
Miscellaneous	637	14.31

Variant Type	Number	% of total
Single-nucleotide substitution	4039	90.76
Single-nucleotide deletion	76	1.71
Single-nucleotide insertion	127	2.85
Same-length multi-nucleotide substitution	12	0.27
Multi-nucleotide deletion	12	0.27
Multi-nucleotide insertion	18	0.40

ASSIGNMENT OF ALLELE NOMENCLATURE

Table 3. Sources of allele nomenclature

Gene Family	URL
Cytochrome P450	http://www.imm.ki.se/CYPalleles
Arylamine N-Acetyltransferase	http://www.louisville.edu/medschool/pharmacology/NAT.html
UDP Glucuronosyltransferase	http://som.flinders.edu.au/FUSA/ClinPharm/UGT

The alleles monitored by locus-specific allele nomenclature committees were collected from the web sites described in Table 3. To ensure correct assignment of allele nomenclature to Celera variants, we implemented a pipeline as illustrated in Fig. 2. Due to very loose representation of data in these allele sites, considerable effort was put into parsing and extracting correct data prior to submitting it to the pipeline for mapping and allele nomenclature assignment. Briefly, the reference sequences of each gene, as reported in the allele web sites, were manually collected, inspected, and modified if necessary. We then calculated the coordinates on reference sequences for each variant from the web sites using the rules of den Dunnen and Antonarakis (ref.3). Any inconsistent results we identified were reported back to the nomenclature committees, and some of them have since been corrected on these allele web sites. Celera variants were mapped to the reference sequences and their coordinates were compared to the variants collected from the allele web sites. Allele nomenclature is assigned when the locations and nucleotide changes of Celera variants both match the ones collected from the allele web sites. Some harder cases, such as those variants adjacent to each others, were manually curated.

Figure 2. Assigning allele nomenclature to DME variants

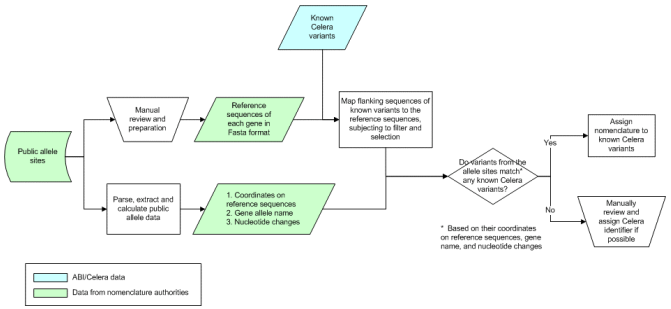
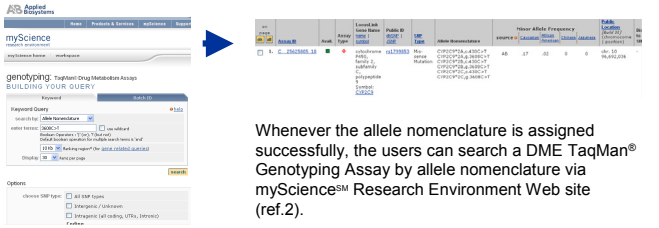


Figure 3. Searching DME assays by allele nomenclature



Whenever the allele nomenclature is assigned successfully, the users can search a DME TaqMan® Genotyping Assay by allele nomenclature via myScience℠ Research Environment Web site (ref.2).

CONCLUSIONS

A comprehensive catalog of >3,000 protein-coding SNPs for 219 DME-related genes was created using the data from a variety of proprietary and public sources, including several prominent allele nomenclature web sites. The results enable the design of a DME genotyping platform for researchers to screen for these polymorphisms.

REFERENCES

1. Levistsk, A., et al. AB Asilomar Conference, September 18-20, 2005, Poster #140
2. The myScience web site URL is http://dme.appliedbiosystems.com
3. den Dunnen, J. T., and Antonarakis, S. E. Human Genetics 109(1): 121-124, 2001

ACKNOWLEDGEMENTS

We thank Dr. Peter Li for guidance and support.

TRADEMARKS/LICENSING