

How Well do the HapMap SNPs “Tag” Functional Variants in Drug Metabolizing Enzyme Genes?

Francisco M. De La Vega¹, Fiona Hyland¹, Katherine Lazaruk¹, Kashif Haque², and Robert A. Welch²

[†]Applied Biosystems, Foster City, CA, and ²Core Genotyping Facility, Division of Cancer Epidemiology and Genetics, SAIC Frederick, National Cancer Institute, Gaithersburg, MD USA

INTRODUCTION

We examined patterns of linkage disequilibrium (LD) among polymorphisms in drug-metabolizing enzyme (DME) genes. DMEs are a class of proteins that are responsible for metabolizing drugs and other xenobiotic compounds. Many of those enzymes show variant alleles among individuals. The variants are often associated with different reactions to drug therapeutic efficacy, adverse reactions and toxicity. We have previously developed and wet-lab validated over 2,000 specific and robust TaqMan[®] drug metabolism genotyping assays for putatively functional polymorphisms (mostly SNPs, but also including multiple nucleotide and insertion/deletions variants) in 220 DME genes. To validate the performance of these assays, we genotyped 180 DNA samples from European, African-American, Chinese and Japanese populations. In addition, we genotyped these DME SNPs on the DNA sample collection of the International HapMap Project obtained from the Coriell Institute. It has previously been noted that rare alleles are less likely to be in high LD with a set of ‘tagging SNPs’, such as those selected using HapMap data. Therefore, we analyzed the LD patterns between HapMap SNPs (release 19) and the putatively functional DME SNPs in our collection on the HapMap samples.

MATERIALS & METHODS

SNP Annotation and Assay Development. DME SNPs were obtained from the literature, DME nomenclature web sites, and by including all non-synonymous variants on DME genes from public databases, Celera SNP discovery, and from the Applera Genomic Initiative gene resequencing project. Flanking sequences of polymorphic sites were mapped to the NCBI build 35 genome assembly. DME polymorphisms were clustered, and functional classification was assigned to the polymorphisms based on their relative positions on DME proteins. Ambiguous results were inspected and resolved manually by experts. Standard allele nomenclature was assigned using a custom bioinformatics pipeline that extracts and calculates the coordinates of DME polymorphisms on the reference sequences designated by the committees. DME alleles were assigned an hCV ID (our internal SNP identifier) if their locations and bases both matched; each represents one non-redundant polymorphic site. Many DME genes are part of large gene families that include several pseudogenes. The resulting large homology is a barrier difficult for most technologies to overcome. Using our extensive bioinformatics pipelines and algorithms, we designed TaqMan SNP Genotyping assays, validated them on the lab by typing a population panel, and repeated design and validation up to 12 times to find the best assay if necessary.

Genotyping. Genotyping of the HapMap Project DNA panel was performed at the Core Genotyping Facility (SAIC Frederick) of the National Cancer Institute, utilizing the protocol specifically recommended by the manufacturer (Applied Biosystems, Foster City, CA, USA) for the TaqMan[®] Drug Metabolism Genotyping Assays. All assays are commercially available. Genotyping quality was outstanding, with a miscall rate among duplicate samples on a plate of only 0.05% and the Mendelian error rate in CEU and YRI of 0%.

Data Analysis. Haploview (v3.2) was used to calculate linkage disequilibrium and select tagging SNPs in batch mode through the use of custom scripts. For the present analysis, we only utilized genotype data from the HapMap Project that passed QC, for SNPs that were polymorphic (MAF >1%) in the CEU panel. We retrieved data for SNPs within and up to 20 Kb up and downstream from the DME gene boundaries. Data from the HapMap SNPs and our DME SNPs was integrated on a per gene basis to perform the study, for 170 DME genes. If a SNP was typed by both us and the HapMap, it was removed from the analysis (i.e. it was considered as tagged by the HapMap by definition). Tagging SNPs were defined as SNPs which tag themselves or another SNP or SNPs with pairwise $r^2 > 0.8$, or with Haplotype $r^2 > 0.8$ in conjunction with one or two additional SNPs. The results shown utilize the aggressive method of the Tagger algorithm as implemented in Haploview; as just very few SNPs were tagged by more than one SNP.

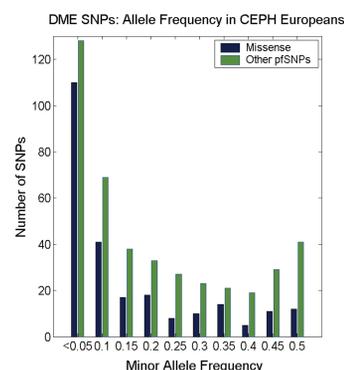
RESULTS

Allele Frequency Spectrum of DME SNPs

Figure 1 shows the correlation between minor allele frequency and coding variation type for the SNPs in our DME collection. The results shown are those obtained during the assay validation on a panel of DNA samples from four populations, and not actually the HapMap samples, though the results are expected to be essentially the same. Some of these monomorphic sites are variants previously reported in the literature, and that are rare, but have a phenotypic impact. Nonsense SNPs have the highest percentage of monomorphic sites, and the lowest minor allele frequency among the polymorphic SNPs, consistent with strong selection. Missense SNPs have a lower allele frequency than other putative functional SNPs (including SNPs in promoter regions and splice sites) in all populations. Also, more DME SNPs are polymorphic in African Americans than in other populations. Polymorphism was observed in 55% of DME SNPs in at least one population 55% of the DME SNPs were polymorphic at least once, this percentage was the same in both HapMap and AB samples.

Figure 1. Allele frequency among DME SNPs related to type of variation.

		SNP Type		
		Missense	Nonsense	Other
# of SNPs		1088	89	1354
CEPH European	# Polymorphic	258	2	502
	Mean MAF	0.151	0.055	0.174
African American	# Polymorphic	361	5	748
	Mean MAF	0.1088	0.054	0.145
Japanese	# Polymorphic	169	4	390
	Mean MAF	0.161	0.155	0.186
Chinese	# Polymorphic	171	3	389
	Mean MAF	0.158	0.1667	0.188



Comparison between HapMap validation and the data collected in this study

Of 1873 putative functional DME SNPs, there were 848 DME SNPs also typed in the HapMap Project, which amount to 45% of the total SNPs in our DME collection (this includes monomorphic sites). Since developing genotyping assays for coding SNPs within highly homologous gene families has proved to be very challenging, this overlap allowed us to analyze the reliability of the HapMap data for such difficult sets of putative functional SNPs.

Table 1. Analysis of data quality for SNPs typed in both HapMap and this study.

Comment	N	%	Reference
Total DME SNPs Analyzed	1873		
Typed also by HapMap	848	45.3%	of DME Collection
Failed HapMap QC	182	21.5%	of HapMap Typed
Discordant QC	107	12.6%	of HapMap Typed
Failed QC at least once	289	34.1%	of HapMap Typed
Passed QC	558	65.8%	of HapMap Typed
QC Pass - Monomorphic	127	22.8%	of QC Pass

Proportion of tagging SNPs that capture DME putative functional variation

Many of the DME SNPs studied here show low heterozygosity as expected for alleles likely to be under strong purifying or directional selection. Rare alleles are less likely to be in high LD with a set

of common ‘tagging SNPs’, such as those selected using HapMap data. We thus investigated how well tagSNPs selected from the combined dataset of the HapMap and DME SNPs can capture the putative functional variants. We were also interested in knowing how many of the DME SNPs not typed by the HapMap project can be tagged by a HapMap tagSNPs. The latter is important since this is an analogous situation to an association mapping study by LD, where common tagSNP selected from the HapMap are intended to capture the information of functional variation on genes in order to report their association with the phenotype under study. We performed our analysis on a per gene basis, where we included HapMap SNPs up to 20 kb up and downstream of the gene boundaries (as defined as the coordinates of the first and last exons and UTRs).

Figure 2. LD Analysis plot for the gene CYP2B6.

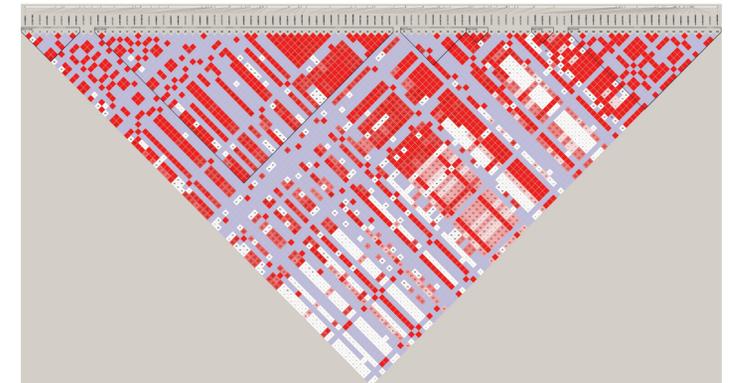


Table 2. Capture of putative functional SNP variation across 170 DME genes by tagging SNPs at $r^2 > 0.8$

SNP Subset	N
Tagged by HapMap	177
Tagged by one or more HapMap SNP	137
Tagging HapMap SNPs	32
Tagged by DME + HapMap SNPs	8
Untagged by any HapMap SNP(s)	108
Tagged by other DME SNP(s)	148
Total tagging SNPs (HapMap + DME)	3330

CONCLUSIONS

To detect phenotypes influenced by putative functional DME SNPs, it will often be necessary to directly genotype the causative variant. Our results suggest that genome-wide and candidate gene/region association studies utilizing subsets of the 3 million validated HapMap SNPs may miss up to 40% of coding functional variation, such as the DME SNPs, in particular rare causative variants. Therefore, investigators interested in DME SNPs, and other putative functional polymorphisms, should use caution in relying solely on tagging SNPs for research on disease association or drug efficacy, toxicity, and metabolism, and should consider typing directly the relevant variants in their studies whenever possible. A good strategy for genetic association studies would be to perform direct association studies that include a comprehensive collection of common putative functional SNPs, or to complement second-stage replication studies with as many as possible non-synonymous coding SNPs. It is also important to consider utilizing a robust genotyping platform that can produce reliable assays for conserved regions such as the TaqMan SNP genotyping assay.

Acknowledgements

We would like to thank Laurie Burdette for providing HapMap annotations for the overlap with DME SNPs. We acknowledge Chia-Chien Chang, Zhiping Gu, Shuang Cai, Rosane Charlab, Alexander Levitsky, and Daniel Ingber for their previous work to retrieve, map, and annotate the DME SNPs. Thanks are also due to Warren Tom and Sueh-Ning Liew for contributions to assay development and validation, and to Tim Harkins for his continuous support of the project.

Legal Notice

For Research Use Only. Not for use in diagnostic procedures. Applied Biosystems is a registered trademark, and AB (Design) and Applera are trademarks, of Applera Corporation or its subsidiaries in the US and/or certain other countries. TaqMan is a registered trademark of Roche Molecular Systems, Inc. All other trademarks are the sole property of their respective owners. © 2006 Applera Corporation. All rights reserved. 135PR02-01