Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

On the challenges of the HapMap resource

Wei Zhang[1] and M. Eileen Dolan[1, 2, 3]

[1]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL 60637, USA; [2]Committee on Clinical Pharmacology and Pharmacogenomics, University of Chicago, Chicago, IL 60637, USA; [3]Cancer Research Center, University of Chicago, Chicago, IL 60637, USA; M. Eileen Dolan* - E-mail: edolan@medicine.bsd.uchicago.edu; Phone: +773-702-4441; Fax: +773-702-0963; * Corresponding author

received December 29, 2007; accepted January 09, 2008; published January 11, 2008

Abstract:
The International HapMap Project provides a key resource of genotypic data on human lymphoblastoid cell lines derived from four major world populations of European, African, Chinese and Japanese ancestry for researchers to associate with various phenotypic data to find genes affecting health, disease and response to drugs. Recently, the HapMap resource has significantly benefited research areas such as gene expression variation studies. Besides some intrinsic limitations, there are a few challenges that should be considered in the next wave of research using this tremendous resource. We suggest that overcoming these challenges or considering the confounding variables in the interpretation of results can provide more insights into the current views of the human genome as well as complex traits such as drug response variation and susceptibility to common diseases.

Keywords: HapMap; lymphoblastoid cell lines; single nuceotide polymorphism; genotype; gene expression

Background:
The goal of the International HapMap Project **[1]** is to develop a haplotype map of the human genome, the HapMap, to describe the common patterns of human DNA sequence variation. Four populations were selected for inclusion in the HapMap: 30 trios from Ibadan, Nigeria (YRI), 30 trios of U.S. residents of northern and western European ancestry (CEU), 45 unrelated individuals from Tokyo, Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB). Epstein Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) were derived from these individuals. The HapMap resource is comprised of genotypic data on ~4 million single nucleotide polymorphisms (SNPs), gene expression data using various microarray platforms, other phenotypic data such as drug response as well as structural variation data (copy number changes) **[2]**. In addition, there is data available regarding cellular sensitivity to various chemotherapeutic agents at PharmGKB **[3]**. Thus, investigators can utilize extensive genotype, gene expression and other phenotypic data on the 270 HapMap samples to perform various genome-wide scans for studies that do not require *a priori* knowledge or experimentation to generate new hypothesis. Because the HapMap samples provide samples from major world populations, investigations into inter-ethnic variation in genotype, gene expression and phenotype are possible. Recently, using the HapMap resource, some exciting progress has been made in the areas of gene expression variation studies and pharmacogenomics **[2]**. Albeit the successes, there are some limitations of the HapMap resource **[2]**. For example, LCLs represent just one human tissue type and gene expression levels vary considerably among tissues. For pharmacologic studies, LCLs may not reflect tumor response or sensitivity of target tissue of known toxicity. Also, studies have suggested that EBV transformation may affect the expression of some genes and certain biological processes in LCLs **[4, 5]**. Therefore, interpretation of results using these cell lines may be biased by this effect. In this paper, we will not delve into these "intrinsic" limitations of the HapMap resource. Instead, we will discuss some of the technical challenges when utilizing the HapMap resource.

Description:

Problem of untyped SNPs

The HapMap resource has allowed whole-genome associations to detect genetic determinants, particularly SNPs that affect gene expression variations (expression quantitative trait loci, eQTLs) and other phenotypes such as drug response **[2]**. However, a statistically significant association between a SNP and a phenotype does not necessarily indicate that the relationship is with a causal variant. The detected SNP could be in linkage disequilibrium (LD) with the casual variant. Therefore, involving the causal SNP or the marker SNP in LD with it will be critical to detect a causal association. With the availability of ~4 million genotypic data on SNPs, the HapMap resource provides a reference catalog of human genetic variations. The problem with the whole-genome associations is, however, whether the HapMap SNPs are sufficient to capture most of the human variation and untyped SNPs. Using the National Institute of

Environmental Health Sciences (NIEHS) Environmental Genome Project SNPs, Tantoso *et al.*, showed that the HapMap SNPs are transferable to the NIEHS SNPs. However, the HapMap SNPs do not capture some SNPs (~45% for CEU and JPT/CHB, ~70% for YRI) **[6]**. Therefore, deep resequencing or a denser genotyping microarray would uncover more SNPs in the HapMap samples so that researchers can be certain that tag-SNPs chosen for association studies are able to comprehensively cover all the variants in the genes. Some efforts to resequence the HapMap samples include 1) the coordination of the HapMap Project and the ENCODE (ENCyclopedia Of DNA Elements) Project**[7]**, 2) the SeattleSNPs Project **[8]** and 3) the NIEHS Environmental Genome Project **[9]**. For example, the SeattleSNPs Project has implemented a large-scale genotyping effort, using Illumina BeadArray technology, to map highly informative tag-SNPs from previously studied SeattleSNPs candidate genes in all unrelated HapMap samples. In contrast, the ENCODE-HapMap Coordination aims to resequence ten 500Kb genomic regions in 48 unrelated individuals (16 YRI, 16 CEU, 8 CHB and 8 JPT) using a PCR-based method. No doubt, the HapMap resource will become an even more powerful tool once integrated with these efforts.

Confounding factors when using the resource

Although exciting results have been generated using the HapMap resource **[2]**, to date few researchers have taken proper care of some potential confounding factors when using these data. We classified these confounding factors into three major groups: 1) non-genetic factors such as batch effect, cell collection time etc. Akey et al. pointed out that the batch effect of a HapMap dataset could significantly impact the interpretation of the results of a gene expression variation study **[10]**; 2) technical confounding factors such as SNPs in probes. The confounding effect of SNPs in probes has been appreciated by researchers **[11]**, however, most available HapMap gene expression datasets using various microarray platforms (eg. Affymetrix Human Focus array, Illumina BeadChip **[2]**) did not consider this effect. A recent version of expression dataset using the Affymetrix Human Exon array took an extra step to filter out probesets affected by SNPs maintained in dbSNP (a database of SNP data curated by the National Center for Biotechnology Information) when summarizing gene expression **[2, 12]**, though a more comprehensive comparison study may be necessary to investigate in detail this effect on gene expression; and 3) other potential confounding factors such as gender and cell proliferation rate need to be considered when interpreting results from particular association studies with expression and or pharmacological response. For example, some drugs are more effective on rapidly growing cells, therefore the proliferation rate may be an intermediary confounding effect. Genes important in cellular susceptibility to a particular drug could include genes that solely affect cellular proliferation.

Conclusion:

The HapMap resource has provided researchers a powerful tool to explore quantitative variation in complex traits such as gene expression and drug response in human populations. Overcoming the challenges associated with this resource or considering the confounding variables in the interpretation of results would facilitate the next wave of research using this tremendous resource and provide more insights into the current views of the human genome as well as complex traits such as drug response variation and susceptibility to common disease.

Acknowledgement:

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group (http://pharmacogenetics.org) study was supported by NIH/NIGMS grant U01 GM61393.

References:

[01] International HapMap Consortium, *Nature*, 426: 789 (2003) [PMID: 14685227]

[02] W. Zhang, *et al.*, *Bioinformatics and Biology Insights*, 2 (2008)

[03] http://www.PharmGKB.org

[04] S. Tobollik, *et al.*, *Blood*, 108: 3859 (2006) [PMID: 16882707]

[05] M. T. Liu, *et al.*, *Oncogene*, 23: 2531 (2004) [PMID: 14716302]

[06] E. Tantoso, *et al.*, *BMC Genomics*, 7: 238 (2006) [PMID: 16982009]

[07] ENCODE Project Consortium, *Science*, 306: 636 (2004) [PMID: 15499007]

[08] http://pga.mbt.washington.edu

[09] http://www.niehs.nih.gov/research/supported/programs/egp/

[10] J. M. Akey, *et al.*, *Nat Genet.*, 39: 807 (2007) [PMID: 17597765]

[11] Y. Gilad, *et al.*, *Genome Res.*, 15: 674 (2005) [PMID: 15867429]

[12] W. Zhang, *et al.*, *Am J Hum Genet.*, (in press)

Edited by P. Kangueane

Citation: Zhang and Dolan, Bioinformation 2(6): 238-239 (2008)