Published in final edited form as:

Curr Bioinform. 2008 September 1; 3(3): 178. doi:10.2174/157489308785909232.

Beyond the HapMap Genotypic Data: Prospects of Deep Resequencing Projects

Wei Zhang¹ and M. Eileen Dolan^{*,1,2,3}

¹ Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, IL 60637, USA

² Committee on Clinical Pharmacology and Pharmacogenomics, The University of Chicago, Chicago, IL 60637, USA

³ Cancer Research Center, The University of Chicago, Chicago, IL 60637, USA

Abstract

The International HapMap Project provides a key resource of genotypic data on human samples including lymphoblastoid cell lines derived from individuals of four major world populations of African, European, Japanese and Chinese ancestry. Researchers have utilized this resource to identify genetic elements that correlate with various phenotypes such as risks of common diseases, individual drug response and gene expression variation. However, recent comparative studies have suggested that the currently available HapMap genotypic data may not capture a substantial proportion of rare or untyped SNPs in these populations, implying that the HapMap SNPs may not be sufficient for comprehensive association studies. In this paper, three large-scale deep resequencing projects covering the HapMap samples: ENCODE (Encyclopedia of DNA Elements), SeattleSNPs and NIEHS (National Institute of Environmental Health Sciences) Environmental Genome Project are discussed. Prospectively, once integrated with the HapMap resource, these efforts will greatly benefit the next wave of association studies and data mining using these cell lines.

Keywords

HapMap; lymphoblastoid cell lines; genotype; single nucleotide polymorphism; resequencing

INTRODUCTION

Unlike Mendelian or rare diseases, the interplay of both genetic and non-genetic factors (e.g. environmental influence) contributes to the basic biology and pathogenesis of common diseases such as heart attack, stroke, Alzheimer's disease, asthma, cancer and diabetes, as well as the individual response to therapeutic agents. A typical example may be the gene-environment (air pollution) interactions in the development of asthma [1,2]. However, genetic factors may still play a major role in the etiology of these diseases. For example, the rapid increase in the population prevalence of type 2 diabetes mellitus (T2DM) in youth can only be explained by changes in lifestyle, however, even when most members of a population have changed their lifestyle, only a minority of children develop diabetes, therefore, genetic factors are important in determining which children are most susceptible [3].

*Address correspondence to this author at the Section of Hematology/Oncology, Department of Medicine, 5841 S Maryland Ave. MC 2115, The University of Chicago, Chicago, IL 60637, USA; Tel: 773-702-4441; Fax: 773-702-0963; edolan@medicine.bsd.uchicago.edu.

Single nucleotide polymorphisms (SNPs), which are DNA sequence variation that occurs when a single nucleotide (A, T, C or G) in the genome sequence is altered, make up about 90% of all human genetic variation [4,5]. Although more than 99% of human DNA sequences are the same between individuals [6,7], variation in DNA sequence can have a major impact on how humans respond to disease and environmental factors. Recent studies have shown that the variation of DNA sequence in the form of SNPs is likely to play a critical role in determining the susceptibility of various common diseases (e.g. Parkinson's disease and heart disease) [8–11] or variability in the response to drugs (eg. anticancer agents and cardiovascular drugs) [12–14], though other types of genotypic variation such as copy number variation (CNV) can also contribute to common disorders [15].

Obtaining a detailed map of genetic variation such as SNPs and CNV in the human genome, therefore, can potentially facilitate studies on the genetic contribution to common diseases or drug response. During the past several years, the International HapMap Project (<http://www.hapmap.org/>) has provided researchers a key resource of genotypic data on a panel of lymphoblastoid cell lines (LCLs) derived from 270 individuals of major world populations in Europe (CEU, Utah, USA residents of northern and western European ancestry), Africa (YRI, Yoruba people from Ibadan, Nigeria) and Asia (JPT, Japanese from Tokyo, Japan; CHB, Han Chinese people from Beijing, China) [16,17] (Fig. 1). The recently released Phase II data of the HapMap Project is comprised of more than 3.1 million SNPs including approximately 25–35% of common SNP variation in the populations surveyed [18]. Besides the SNPs, other types of genotypic (e.g. CNV) and phenotypic data (eg. drug response, gene expression) have been made public by the HapMap Project and other laboratories. We previously reviewed in detail the most recent HapMap resource available to researchers [19]. The progress of the HapMap Project has allowed whole-genome association studies to identify the genetic determinants responsible for phenotypic variations [20–24]. Particularly, exciting progress has been made in studying gene expression variation within and between human populations (gene expression studies) and the identification of genetic determinants responsible for cytotoxicity to anti-cancer drugs (pharmacogenomic studies) [21–26] (for a review see Zhang *et al.* [19]). In contrast to the traditional candidate-gene approaches, the availability of the HapMap resource allows for whole-genome association studies that do not require *a priori* knowledge. This flexibility is of considerable value in the identification and exploration of novel therapeutic targets or genetic factors that are related to common complex traits.

In spite of the successes in the studies of gene expression and pharmacogenomics [19], there are some challenges that need particular attention when utilizing this tremendous resource in the next wave of analyses [27]. Some of the challenges are “intrinsic”. In other words, these are limitations of the HapMap resource that can not be easily improved upon with future progress. One example is that the HapMap samples only represent one cell type (i.e. lymphoblast), while researchers already know that some pharmacology-related genes (e.g. *UGT1A*) have tissue-specific expression patterns [28]. Two potential confounding factors in the use of these cell lines is the inter-ethnic differences in cell collection time (CEU samples collected many years earlier than other populations) and inter-individual differences in rate of cellular proliferation [27]. Notably, Akey *et al.* demonstrated that the cell collection time could be responsible for the observed differences in gene expression between the HapMap populations of Asian and European ancestry in one recent study [29]. Besides the “intrinsic” challenges, one major concern in utilizing the HapMap resource is the inability of the HapMap data to capture rare or untyped SNPs [27]. In whole-genome association studies, 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 causal variant. Therefore, involving the causal SNP or the marker SNP in LD will be critical in detecting a causal association. With the current version of Phase II HapMap genotypic data comprised of more than 3.1 million SNPs [18], the coverage of the

HapMap data is believed to be generally high, though a couple of studies have shown that the HapMap genotypic data can not capture a substantial proportion of untyped SNPs [30–32], suggesting that deep resequencing may be needed to uncover more SNPs in the missing regions. In this review, three large-scale deep resequencing projects covering the HapMap samples: ENCODE (Encyclopedia of DNA Elements) [33], SeattleSNPs and NIEHS (National Institute of Environmental Health Sciences) Environmental Genome Project [34] are discussed. Comparative studies between the SNP coverage of these resequencing projects and the HapMap genotypic data are also reviewed. Prospectively, once integrated with the HapMap resource, these resequencing efforts will greatly benefit the next wave of association studies and data mining efforts using the Hap-Map cell lines.

HAPMAP-ENCODE COORDINATION

The ENCODE Project launched by the NHGRI (National Human Genome Research Institute) aims to provide a more biologically informative representation of the human genome by using high-throughput methods to identify and catalogue the functional elements it encodes [33]. The project is being conducted in three phases: a pilot project phase, a technology development phase and a planned production phase. The pilot phase is testing and comparing existing methods to rigorously analyze a defined portion of the human genomic sequence. The goal of the first two phases of the ENCODE project is to identify a suite of approaches that will allow the comprehensive identification of all the functional elements in the human genome. Through the ENCODE pilot phase, the NHGRI expects to assess the abilities of different approaches to be scaled up in an effort to analyze the entire human genome and to find gaps in our ability to identify functional elements in genomic sequence (<http://www.genome.gov/10005107/>). Results from the pilot phase of the ENCODE Project include the generation and analysis of functional data from multiple, diverse experimental approaches performed on a targeted 1% of the human genome (44 genomic regions, ~30 Mb) [35]. A variety of sequence elements including genes, promoters, enhancers, repressor or silencer sequences, exons, replication origin and termination sites, transcription factor binding sites, methylation sites, DNase I hypersensitive sites, chromatin modifications, conserved sequences, and RNA transcripts have been considered in this phase [35]. The set of ENCODE targets was chosen to represent a range of genomic features such as the presence of well-studied genes and the existence of a substantial amount of comparative sequence data [33,35]. By the conclusion of the pilot phase, the 44 ENCODE targets will be considered the most well-characterized regions in the human genome and will likely be the basis of many future genomic studies. Beyond the pilot phase is the scale-up of the ENCODE Project. This extension will result in increasing the data set in multiple dimensions: not only will the full genome landscape be studied, but also the range of the study will expand by increasing the number of factors and modifications, as well as the number of tissues and states (eg. normal vs. affected) to be considered [36].

The International HapMap Project is already coordinating their efforts to ensure effective synergy with ENCODE activities. The ENCODE Project adds to the HapMap genotypic data by addressing many more types of sequence-based function further connecting the genetic code to the mechanistic biochemical understanding of the human genome [36]. Currently, the HapMap-ENCODE Coordination (<http://www.hapmap.org/downloads/encode1.html.en>, accessed on Feb. 15, 2008) aims to produce a dense set of genotypes across large genomic regions. Ten 500-kb regions of the human genome were resequenced in 48 unrelated HapMap LCL samples (16 YRI, 8 JPT, 8 CHB, and 16 CEU samples) using a PCR-based method. Approximately 20,000 SNPs were identified in the 10 HapMap-ENCODE regions. Some of these were already represented in dbSNP [37,38] (a public SNP database maintained by the National Center for Bio-technology Information, NCBI), while others were discovered during the resequencing. All SNPs identified, along with SNPs in dbSNP, were genotyped in the 269 HapMap LCL samples (90 YRI, 44 JPT, 45 CHB and 90 CEU samples). In addition, Perlegen

Sciences, Inc. (Mountain View, CA) will genotype all SNPs in the remaining 34 ENCODE regions in all of the HapMap DNA samples as part of its genotyping for the HapMap Project. This study will provide dense genotypic data to allow the development and assessment of methods of analysis.

SEATTLESNPS PROJECT

The SeattleSNPs Project (<http://pga.gs.washington.edu/>) is funded as part of the NHLBI (National Heart Lung and Blood Institute) Programs for Genomic Applications. The SeattleSNPs Project is focused on identifying, genotyping, and modeling the associations between SNPs in candidate genes and pathways that underlie inflammatory responses in humans. Besides samples from Hispanic and African-American populations, the SeattleSNPs Project is now generating comprehensive SNP data for HapMap derived LCL samples from both the European and African populations. The SeattleSNPs Project has implemented a large-scale genotyping effort, using the Illumina BeadArray technology, to map highly informative tag-SNPs from previously studied SeattleSNPs candidate genes in all unrelated HapMap samples (60 CEU, 60 YRI, 45 CHB, and 45 JPT samples). Currently, 247 genes have been genotyped with 8 more in progress (http://pga.gs.washington.edu/summary_stats.html, accessed on Feb. 15, 2008). Because the SeattleSNPs Project is focused on genes related to inflammatory diseases, only the genes in these particular pathways such as immune response are being deep resequenced. Therefore, this resource will be useful to complement with the HapMap genotypic data for a specific set of target genes.

NIEHS ENVIRONMENTAL GENOME PROJECT

Many diseases are the outcome of a complex interrelationship between multiple genetic and environmental factors. Research suggests that individual susceptibility is influenced more by certain genes than by exposure to environmental agents. To better understand how individuals differ in their susceptibility to environmental agents and how these susceptibilities change over time, the NIEHS developed the Environmental Genome Project (EGP) (<http://egp.gs.washington.edu/>) in 1997 [34]. The effort of the NIEHS Environmental Genome Project is focused on examining the relationships between environmental exposures, interindividual sequence variation in human genes and disease risk in U.S. populations [34]. The NIEHS/EGP identified a group of human genes that are likely to influence the outcome of environmental exposures. Polymorphic variants in these “environmentally responsive” genes are being identified by sequencing in a predefined set of human DNA samples representing the diversity of the United States. The first phase of the effort is focused on finding SNPs in human genes involved in DNA repair and cell cycle pathways (<http://www.niehs.nih.gov/research/supported/programs/egp/>).

Technically, automated DNA sequencing is being used to identify and genotype SNPs in human candidate genes. All SNPs have been identified using only high quality sequence data and each SNP reported from the NIEHS/EGP SNPs program has been confirmed by multiple individual laboratories and/or in multiple reactions. The NIEHS/EGP SNPs Project is now generating comprehensive SNP data for HapMap derived samples from the European, Asian and African populations. Similar to the SeattleSNPs Project, they also implemented a large-scale genotyping effort, using the Illumina BeadArray technology, to map highly informative tag-SNPs from previously studied NIEHS/EGP candidate genes in the HapMap LCL samples (60 CEU, 60 YRI, 45 CHB, and 45 JPT samples). So far, 624 genes have been genotyped with 4 additional genes in progress (http://egp.gs.washington.edu/finished_genes.html, accessed on Feb. 15, 2008). Because the NIEHS/EGP is focused on “environmentally responsive” genes, only genes in pathways such as apoptosis and DNA repair are being deep resequenced.

Therefore, similar to the SeattleSNPs resource, the data will complement the HapMap genotypic data only if the gene of interest is in their list of resequencing targets.

HOW WELL DO THE HAPMAP GENOTYPIC DATA CAPTURE UNTYPED SNPS?

The coverage of human genomic variations will substantially affect the success of genome-wide association studies. Therefore, one of the major challenges of the HapMap resource is the significant number of untyped SNPs [27]. The HapMap resource has allowed whole-genome associations to detect common genetic variants, particularly SNPs that affect gene expression variation (expression quantitative trait loci, eQTLs) and other phenotypes such as drug response [19]. With the availability of more than 3.1 million genotypic data on SNPs, the HapMap resource provides a reference catalog of human genetic variation. The problem with the whole-genome associations is, however, whether the HapMap SNPs are sufficient to capture most of the human variation particularly with the number of including untyped SNPs. In theory, progress in the deep resequencing projects such as ENO-CDE, SeattleSNPs and NIEHS/EGP using the same HapMap LCL samples would allow researchers to estimate the level of untyped SNPs in the HapMap resource. Recently, a couple of studies compared the coverage of the HapMap genotypic data and deep resequencing projects [30–32]. The results from these studies shed some light into the problem of untyped SNPs in the HapMap resource.

Using the NIEHS/EGP SNPs, Tantoso *et al.* analyzed the coverage of the HapMap genotypic data [30]. They first determined that the HapMap genotypic data were transferable to the NIEHS/EGP SNPs. However, their analysis suggested that the HapMap data were not robust enough to capture the untyped variants for most of the human genes they surveyed. The performance of SNPs in the European and Asian samples were marginal in capturing the untyped variants, i.e. approximately 55%. In contrast, the SNPs from the HapMap YRI samples only capture approximately 30% of the variants. Therefore, the HapMap SNPs do not capture a substantial proportion of untyped SNPs (~45% for CEU and JPT/CHB, ~70% for YRI). Fig. (2) shows an example (*BRAC1*) for the SNP coverage comparison between NIEHS/EGP and HapMap. Based on these results, they concluded that resequencing might be necessary to uncover more untyped SNPs in the missing region.

By resequencing 334 genes in the NIEHS/EGP, Taylor *et al.* were able to compare haplotypes of the resequenced genes with the HapMap genotypic data [31]. Approximately 15% of common SNPs in the NIEHS/EGP genes were found to have genotypic data in the HapMap database. They further resampled the NIEHS/EGP data to examine the effect of increasing the number of HapMap SNPs to 5 million, and estimated that approximately 40% of common SNPs in genes would be sampled and that half of the genes would have sufficient SNPs to identify all common haplotypes. Their data suggested that this inability to distinguish common haplotypes of genes might result in loss of power when examining haplotype-disease association.

Independent of the deep resequencing projects reviewed above, Takeuchi *et al.* assessed the SNP coverage in the HapMap database for a total of 1,304 subjects from the Japanese population by combining resequencing and high-density genotyping approaches. Of the 738 common SNPs that they detected by resequencing, 58% had already been genotyped in the HapMap, and 31% were not genotyped but had a proxy SNP in the HapMap with a LD coefficient $r^2 \geq 0.8$, whereas 11% were not represented in the current Hap-Map database [32]. They concluded that the HapMap coverage appeared to be high although not thorough for SNPs in the Japanese HapMap samples.

It is noteworthy, however, that these studies did not compare the whole genome. Only a small fraction of the human genome was compared between the HapMap genotypic data and resequencing results. This was largely due to the incomplete coverage of the currently available resequencing data, especially the fact that projects such as NIEHS/EGP cover only a predefined set of genes. With the progress of the whole-genome resequencing efforts such as the ENCODE Project [33], a more comprehensive estimation would be possible in the foreseeable future. On the other hand, these comparative studies used earlier versions of the HapMap database (e.g. release 16 for Taylor *et al.* [31]), therefore, a newer version could have better coverage if used. Still, these studies clearly demonstrated the value of deep resequencing projects in uncovering untyped SNPs that were not being captured by the current HapMap genotypic dataset.

CONCLUSIONS AND OUTLOOK

The HapMap resource has provided researchers a powerful tool to explore quantitative variation in complex traits [19]. Although the current HapMap genotypic data is comprised of more than 3.1 million common SNPs, it remains a challenging task to identify rare or untyped SNPs using this resource [27]. Comparative studies using resequencing data have shown that the HapMap coverage is generally high though not robust enough for a substantial proportion of untyped SNPs [30–32]. Therefore, deep resequencing projects such as those described here will be useful in uncovering more SNPs than the currently available genotypic data. As a matter of fact, investigators using these resequencing data have demonstrated the values of these projects in association studies for the search of disease-related genetic determinants. For example, using the NIEHS/EGP whole-gene resequencing data, Cox *et al.* selected haplotype tagging SNPs and identified a haplotype in the *BRCA1* gene that was associated with an approximately 20% increase in risk of sporadic breast cancer in the general population [39]. In contrast, the SeattleSNPs Project allowed the identification of a common *VLDLR* polymorphism that interacts with *APOE* genotype in the prediction of carotid artery disease risk [40]. These deep resequencing projects will facilitate the next wave of research efforts using the HapMap resource and provide more insights into the current views of the human genome as well as studies relating genetic variation with complex traits such as drug response variation and susceptibility to common diseases.

Acknowledgments

The authors wish to thank other members in the Dolan lab for helpful discussion that led to this review. This Pharmacogenetics of Anticancer Agents Research (PAAR) Group (<http://pharmacogenetics.org>) work was supported by the NIH/NIGMS grant U01 GM61393.

References

1. McLeish S, Turner SW. Gene-environment interactions in asthma. *Arch Dis Child* 2007;92:1032–5. [PubMed: 17954484]
2. London SJ. Gene-air pollution interactions in asthma. *Proc Am Thorac Soc* 2007;4:217–20. [PubMed: 17607002]
3. Gill-Carey O, Hattersley AT. Genetics and type 2 diabetes in youth. *Pediatr Diabetes* 2007;8 (Suppl 9):42–7. [PubMed: 17991132]
4. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291:1304–51. [PubMed: 11181995]
5. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. [PubMed: 11237011]
6. Altshuler D, Pollara VJ, Cowles CR, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 2000;407:513–6. [PubMed: 11029002]

7. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–33. [PubMed: 11237013]
8. Mayeux R. Mapping the new frontier: complex genetic disorders. *J Clin Invest* 2005;115:1404–7. [PubMed: 15931374]
9. Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet* 2000;9:2403–8. [PubMed: 11005795]
10. Kullo IJ, Ding K. Mechanisms of disease: The genetic basis of coronary heart disease. *Nat Clin Pract Cardiovasc Med* 2007;4:558–69. [PubMed: 17893684]
11. Tan EK. The role of common genetic risk variants in Parkinson disease. *Clin Genet* 2007;72:387–93. [PubMed: 17868389]
12. Johnson JA, Cavallari LH. Cardiovascular pharmacogenomics. *Exp Physiol* 2005;90:283–9. [PubMed: 15778411]
13. Nagasubramanian R, Innocenti F, Ratain MJ. Pharmacogenetics in cancer treatment. *Annu Rev Med* 2003;54:437–52. [PubMed: 12525681]
14. Efferth T, Volm M. Pharmacogenetics for individualized cancer chemotherapy. *Pharmacol Ther* 2005;107:155–76. [PubMed: 15890408]
15. Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet* 2007;3:1787–99. [PubMed: 17953491]
16. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437:1299–320. [PubMed: 16255080]
17. International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–96. [PubMed: 14685227]
18. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–61. [PubMed: 17943122]
19. Zhang W, Ratain MJ, Dolan ME. The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform Biol Insights* 2008;2:15–23. [PubMed: 18392109]
20. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet* 2007;39:1217–24. [PubMed: 17873874]
21. Storey JD, Madeoy J, Strout JL, et al. Gene-expression variation within and among human populations. *Am J Hum Genet* 2007;80:502–9. [PubMed: 17273971]
22. Spielman RS, Bastone LA, Burdick JT, et al. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet* 2007;39:226–31. [PubMed: 17206142]
23. Huang RS, Duan S, Bleibel WK, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci USA* 2007;104:9758–63. [PubMed: 17537913]
24. Huang RS, Duan S, Shukla SJ, et al. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet* 2007;81:427–37. [PubMed: 17701890]
25. Zhang W, Duan S, Kistner EO, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet* 2008;82:631–40. [PubMed: 18313023]
26. Duan S, Huang RS, Zhang W, et al. Genetic architecture of transcript-level variation in humans. *Am J Hum Genet* 2008;82(5):101–13.
27. Zhang W, Dolan ME. On the challenges of the HapMap resource. *Bioinformatics* 2008;2:238–9. [PubMed: 18317571]
28. Zhang W, Liu W, Innocenti F, Ratain MJ. Searching for tissue-specific expression pattern-linked nucleotides of UGT1A isoforms. *PLoS ONE* 2007;2:e396. [PubMed: 17460763]
29. Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. *Nat Genet* 2007;39:807–8. [PubMed: 17597765]
30. Tantoso E, Yang Y, Li KB. How well do HapMap SNPs capture the untyped SNPs? *BMC Genomics* 2006;7:238. [PubMed: 16982009]

31. Taylor JA, Xu ZL, Kaplan NL, Morris RW. How well do HapMap haplotypes identify common haplotypes of genes? A comparison with haplotypes of 334 genes resequenced in the environmental genome project. *Cancer Epidemiol Biomarkers Prev* 2006;15:133–7. [PubMed: 16434598]
32. Takeuchi F, Serizawa M, Kato N. HapMap coverage for SNPs in the Japanese population. *J Hum Genet* 2008;53:96–9. [PubMed: 18043864]
33. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306:636–40. [PubMed: 15499007]
34. Livingston RJ, von Niederhausern A, Jegga AG, et al. Pattern of sequence variation across 213 environmental response genes. *Genome Res* 2004;14:1821–31. [PubMed: 15364900]
35. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799–816. [PubMed: 17571346]
36. Weinstock GM. ENCODE: more genomic empowerment. *Genome Res* 2007;17:667–8. [PubMed: 17567987]
37. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000;28:352–5. [PubMed: 10592272]
38. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–11. [PubMed: 11125122]
39. Cox DG, Kraft P, Hankinson SE, Hunter DJ. Haplotype analysis of common variants in the BRCA1 gene and risk of sporadic breast cancer. *Breast Cancer Res* 2005;7:R171–5. [PubMed: 15743496]
40. Crawford DC, Nord AS, Badzioch MD, et al. A common VLDLR polymorphism interacts with APOE genotype in the prediction of carotid artery disease risk. *J Lipid Res* 2008;49:588–96. [PubMed: 18056683]
41. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006. [PubMed: 12045153]

Fig. (1).

Lymphoblastoid cell line samples from the International HapMap Project are derived from individuals of European (30 parents-child trios), African (30 parents-child trios) and Asian (45 unrelated Japanese and 45 unrelated Han Chinese) ancestry. CEU: Centre d'Etude du Polymorphisme Humain (CEPH) samples from residents of northern and western European ancestry in Utah, USA; YRI: Yoruba people from Ibadan, Nigeria; JPT: Japanese people from Tokyo, Japan; CHB: Han Chinese people from Beijing, China. The map template was downloaded at <http://english.freemap.jp/> under the Creative Commons Attribution 3.0 License.

Fig. (2).

UCSC Genome Browser (<http://genome.ucsc.edu/>) [41] view of the SNP coverage comparison between NIEHS/EGP and HapMap. Gene *BRAC1* (chr17:38,449,024-38,531,858, Human March 2006 Assembly) has been resequenced in the unrelated HapMap samples by NIEHS/EGP.