Published in final edited form as:

Pharmacogenomics. 2010 February ; 11(2): 249–256. doi:10.2217/pgs.09.173.

Impact of the 1000 Genomes Project on the next wave of pharmacogenomic discovery

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

Wei Zhang: ; M Eileen Dolan: edolan@medicine.bsd.uchicago.edu

¹ Section of Hematology/Oncology, Department of Medicine, 900 East 57th Street, KCBD Room 7100, The University of Chicago, Chicago, IL 60637, USA, Tel.: +1 773 702 4441 Fax: +1 773 702 0963

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

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

Abstract

The 1000 Genomes Project aims to provide detailed genetic variation data on over 1000 genomes from worldwide populations using the next-generation sequencing technologies. Some of the samples utilized for the 1000 Genomes Project are the International HapMap samples that are composed of lymphoblastoid cell lines derived from individuals of different world populations. These same samples have been used in pharmacogenomic discovery and validation. For example, a cell-based, genome-wide approach using the HapMap samples has been used to identify pharmacogenomic loci associated with chemotherapeutic-induced cytotoxicity with the goal to identify genetic markers for clinical evaluation. Although the coverage of the current HapMap data is generally high, the detailed map of human genetic variation promised by the 1000 Genomes Project will allow a more in-depth analysis of the contribution of genetic variation to drug response. Future studies utilizing this new resource may greatly enhance our understanding of the genetic basis of drug response and other complex traits (e.g., gene expression), therefore, help advance personalized medicine.

Keywords

drug response; genetic variation; HapMap; lymphoblastoid cell lines; pharmacogenomics; SNP

Although environment, diet, age, gender and lifestyle as well as other nongenetic factors (e.g., socioeconomic status) can influence a patient's response to therapeutic treatments, understanding an individual's genetic makeup is believed to be the key to realizing personalized medicine that aims to maximize drug efficacy and minimize adverse side effects. Personalized medicine is particularly appealing for oncologists because of the severity of adverse events and the high likelihood of mortality associated with nonresponse. Side effects can include nephrotoxicity, neurotoxicity and ototoxicity [1,2], and this toxicity is often observed earlier

Correspondence to: M Eileen Dolan, edolan@medicine.bsd.uchicago.edu.

For reprint orders, please contact: reprints@futuremedicine.com

Financial & competing interests disclosure

Some of the research described in this paper was funded by the Pharmacogenetics of Anticancer Agents Research (PAAR) Group (www.pharmacogenetics.org) grant NIH/NIGMS U01GM61393, the University of Chicago Breast Cancer SPOR grant NIH/NCI P50CA125183 and NCI CA136765. The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

than the therapeutic effect [3]. The fundamental problem is that many anticancer drugs present a narrow therapeutic index; thus, small changes in dosage can cause unacceptable toxic responses that, in extreme cases, could lead to fatalities. Pharmacogenomics holds the promise to advance individualized medicine so that therapeutic decision-making would be based on each individual patient's own genetic makeup.

Similar to complex health-related phenotypes, such as the risks of some common diseases (e.g., cancer, diabetes, heart disease and stroke), drug response is often affected by multiple genes in addition to other nongenetic factors. Genes responsible for drug intake, drug metabolism and drug excretion can be involved in an individual's response to therapeutic agents. Previous studies on some well-characterized candidate genes (e.g., drug-metabolizing enzymes) have suggested that DNA variations, especially those in the form of SNPs, in genes that code for these enzymes can influence their ability to break down, convert and efficiently eliminate drugs from the body. A recent example is the observation that reduced cytochrome P450 2D6 activity leads to the therapeutic failure of tamoxifen in the prevention and treatment of breast cancer, as a result of absence of conversion of the prodrug to its active forms [4]. Some classical examples include the identification of genetic polymorphisms in thiopurine S-methyltransferase (*TPMT*), which lead to decreased *TPMT* enzyme activity and subsequently increased 6-mercaptopurine toxicity [5]; decreased activity of *UGT1A1**28 polymorphism is associated with an increased risk of irinotecan treatment-associated neutropenia [6]. With the launch of the Human Genome Project and several parallel research efforts, including the International HapMap Project, which aimed to develop a haplotype map of the human genome to describe the common patterns of human DNA sequence variation [7,8,101], it is now possible to scan the entire human genome or targeted genomic regions to identify genetic determinants responsible for drug-induced effects. Unlike the traditional candidate-gene approach, genome-wide association studies (GWAS) do not require *a priori* assumptions and, therefore, apply an unbiased approach. Notably, pharmacogenomic approaches have been applied to identifying candidate loci associated with response to therapeutic treatments for various diseases (e.g., asthma [9], psychiatric disorders [10], leukemia [11,12], stroke [13] and cardiovascular disease [14]). Particularly, during the past few years, GWAS using the Epstein Barr Virus-transformed lymphoblastoid cell lines (LCLs; e.g., the HapMap samples) have demonstrated the feasibility of integrating whole genome gene expression [15,16] and genotypic data (e.g., >3.1 million SNPs [17]) to identify genes and/or genetic variants associated with the cytotoxicities of anticancer agents, for example 5-fluorouracil [18,19], docetaxel [19], etoposide [20], cisplatin [21], daunorubicin [22], carboplatin [23], cytarabine [24–26] and gemcitabine [25,26] (for reviews see [27–29]). Although the current LCL-based model has been demonstrated to be useful in pharmacogenomic discovery, we will provide our perspective on how the ongoing 1000 Genomes Project, which will generate a much more detailed map of genetic variation using over 1000 LCL samples from diverse populations [102], could help overcome some of its limitations (e.g., untyped SNPs in the HapMap samples) and, therefore, benefit the next wave of pharmacogenomic discovery.

Challenges & concerns of the current studies

The LCLs, particularly the HapMap samples, have proved to be a very useful model for pharmacogenomic discovery, which may represent the best model for hematologic toxicities associated with chemotherapeutic agents [28,29]. One of the major advantages of this cell-based model is the avoidance of giving chemotherapy to unaffected family members for genetic studies and the enormous amount of publicly available genotypic (e.g., SNPs) and phenotypic (e.g., gene expression) data on these samples (National Institute of General Medical Sciences Cell Repository [103]). Another major advantage of this model is that the HapMap samples were derived from three major geographical populations (Caucasian residents with northern and western European ancestry from UT, USA [CEU]; Yoruba people in Ibadan, Nigeria

[YRI]; and Asian samples, including Japanese in Tokyo [JPT] and Han Chinese in Beijing, China [CHB]), thereby allowing for inter-ethnic comparisons in cellular sensitivity to drugs. In fact, previous studies have shown significant differences in cytotoxicities to certain anticancer drugs between human populations [30].

However, there are currently some important limitations and challenges associated with the HapMap LCL samples [31]. Some limitations are ‘intrinsic’. For example, the LCLs represent only one tissue type from apparently healthy individuals; therefore, they may not reflect tumor response or sensitivity of target tissue of known toxicity. In addition, only approximately 50–60% of human genes are estimated to be expressed in LCLs [32]. Obviously, a more comprehensive understanding of the genetic basis of drug response will need to consider other tissues or possibly tumors.

Fortunately, some of the limitations and challenges of the current model will be addressed with advancement of technologies, development of new algorithms and better study design. Since the original HapMap panel is comprised of 90 CEU (30 parents–offspring trios), 90 YRI (30 parents–offspring trios), 45 unrelated CHB and 45 unrelated JPT samples, there may not be enough statistical power to identify genetic variants with small-to-medium effect sizes. A recent study showed that there exists quite significant genetic variation between the two Asian samples (CHB and JPT) [33], suggesting that simply combining these samples as a single Asian population in studies might lead to spurious associations. This limitation may especially exacerbate the power issue for the Asian samples (45 samples each population), because over 55 samples would be needed to attain the power of 80% to identify a variant with medium effect size (e.g., 0.15). On the other hand, although the coverage of the HapMap Project data is believed to be generally high, comparison studies have shown that the HapMap genotypic data may not be able to capture a substantial proportion of untyped SNPs [34,35]. For example, Tantoso *et al.* showed that the SNPs from the HapMap YRI samples capture only approximately 30% of the variants [35], when compared with a deep-resequencing project from the NIEHS Environmental Genome Project [104], and overall, the HapMap SNPs were not robust enough to capture the untyped variants for most of the genes they surveyed [35]. This is not surprising, because the efforts of the International HapMap Project have been focused on characterizing common genetic variants with allele frequencies of greater than 5% [7,8,101]. Thus, for example, untyped or unknown rarer SNPs with large effects cannot be identified using the currently available data on these samples. In addition, although the HapMap LCL model is of tremendous value in the discovery stage, before being tested in clinical trials, these identified pharmacogenomic loci will need to be thoroughly validated in independent replication sets and/or their functions may need to be determined. This raises a challenge to the current HapMap LCL model (i.e., sample coverage of only three major populations), which, although allowing limited cross-validation between the three populations, does not accommodate validations either in the same population or across more geographical populations. Furthermore, the associated loci from the current studies may simply be proxies for underlying causal genetic variants, which may not be genotyped in the HapMap data. Theoretically, these ‘nonintrinsic’ limitations or challenges could be overcome by, for example, a large-scale deep-resequencing project that aims to provide a much more detailed map of human genetic variation on a much larger number of world-wide samples. The 1000 Genome Project aims to do just that.

Next-generation

sequencing technologies—The success of large-scale sequencing efforts depends on the capability to process a large number of samples in parallel and obtain reliable sequencing data in an acceptable turnaround time and at a sustainable cost. Recently, several new sequencing instruments referred to as ‘next-generation’ or ‘massively parallel’ sequencing platforms have become available for the fast, inexpensive sequencing of whole genomes [36,37]. Some

relatively mature platforms include the GS-FLX™ (454) sequencer (Roche, CT, USA), the Genome Analyzer (Illumina, CA, USA) and the Sequencing by Oligo Ligation and Detection (SOLiD™; Applied Biosystems, CA, USA), as well as the so-called single-molecule sequencing technologies [38,39] from Helicos Biosciences (MA, USA) and Pacific Biosciences (CA, USA). In contrast to conventional capillary-based sequencing, these next-generation sequencers are able to process millions of sequencing reads in parallel rather than 96 at a time, although individual platforms have different performance characteristics.

The 1000 Genomes Project

Using the next-generation sequencing technologies (i.e., the Illumina platform and the Applied Biosystems SOLiD) [36], a deep-resequencing project launched in 2008, the 1000 Genomes Project, ambitiously aims to provide the most detailed map of human genetic variation yet through genotyping at least 1000 human genomes from world-wide populations [102]. As the 1000 Genomes Project focuses on samples for which consent has been obtained for open access on the web without needing approval for each use, these requirements have led to the choice of the HapMap samples (HapMap Phase 3 panel) [102]. Besides the original samples from the International HapMap Project (CEU, YRI, JPT and CHB) [7,8,101], the following seven populations are also included in the study: Luhya in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); Toscani in Italy (TSI); Gujarati Indians in Houston, TX, USA (GIH); Chinese in Metropolitan Denver, CO, USA (CHD); people of Mexican ancestry in LA, California, USA (MEX); and people of African ancestry in the southwestern USA (ASW). Particularly, the Epstein Barr Virus-transformed LCLs and DNA samples derived from these individuals are available through the NIGMS Human Genetic Cell Repository (for the CEU samples) [103] and the NHGRI Sample Repository for Human Genetic Research (for the other ten populations) [105]. The specified aims of this project are to identify over 95% of the variants with allele frequencies of more than 1% in parts of the human genome that can be sequenced, over 95% of the variants with allele frequencies over 0.1–0.5% in exons, as well as structural variants, such as copy-number variants (CNVs), other insertions and deletions, and inversions, including sequence-level understanding of breakpoints [106].

In December 2008, the 1000 Genomes Project announced the release of the first set of SNP calls for four individuals (three samples from a CEU parents–child trio and one YRI sample) that are part of the high-coverage pilot project ($>20\times$) [102]. In addition to these four samples, the current data release (accessed on 28 July 2009) covers more than 700 samples of the low-coverage project ($2\times$). In addition to SNP genotypic calls, for each LCL sample, the raw project data [102], including FASTQ files (nucleotides and quality assessments), Binary Simple Alignment/Map files, and FASTA files for the human genome reference assembly, have also been released through the Short Read Archive at the NCBI [107]. The NCBI Short Read Archive is specifically designed for short read data, and will be making the complete project data available in the future. Currently, indel calls are available for the trio children NA12878 (CEU) and NA19240 (YRI; May, 2009). Additional updates of the 1000 Genomes Project data are expected to be released regularly.

Browsing the 1000 Genomes data

Owing to the huge amount of data and the new data types, the analysis of the 1000 Genomes Project data poses formidable informatics challenges. The 1000 Genomes Project provides a web-based browser to facilitate immediate analysis of the 1000 Genomes data (December 2008) by the whole scientific community [108]. This Ensembl-based browser integrates the SNP calls and read coverage for the four genomes (three CEU and one YRI) in the high-coverage pilot 2. The current version (accessed 7 October 2009) supports the viewing of the consequences of sequence variation at the level of each transcript in the genome (Transcript

SNP View) as well as showing read-depth data alongside SNPs (SeqAlign View) relative to the NCBI Build 36 reference (October 2005). Other bioinformatics tools, such as EagleView [40,109] and MapView [41,110], which were tailored for the next-generation sequencing technologies, would also be useful for visualizing and analyzing these new data.

In addition, a pharmacogene database enhanced by the 1000 Genomes Project was built for the community to immediately evaluate and utilize these newly released data [42,111]. Particularly, this database can be used to access SNP genotypic calls (both novel SNPs and known SNPs based on the dbSNP v129) of 39 pharmacogenetic candidate genes, maintained by the Very Important Pharmacogenes (VIP) project of the Pharmacogenetics Knowledge Base (PharmGKB) on 35 HapMap CEU and 26 HapMap YRI samples (April 2009) [43,112]. The VIP project is an initiative to provide annotated information regarding genes, variants, haplotypes and splice variants of particular relevance for pharmacogenetics and pharmacogenomics. A major advantage of this pharmacogene database [42,111] is that it allows the convenient extraction of genotypic calls on novel SNPs that have not been genotyped in the previous HapMap Phase 2 data [17,101]. Table 1 shows the summary of identified novel (i.e., SNPs not recorded in dbSNP v129) and known SNPs (i.e., SNPs included in dbSNP v129) in the 21 VIPs expressed in the CEU samples (based on criteria in Zhang *et al.* [15]). It is clear from Table 1 that there exists a substantial number of novel SNPs in many of these candidate genes (e.g., *AHR* has 25 novel SNPs vs 29 known SNPs). Convenient links to resources, such as the PharmGKB web-site, the Database of Genomic Variants [44], Gene Ontology [45] and the SNP and CNV Annotation Database (SCAN) [46,113], are also provided to allow researchers to access important and relevant information on these genes. Even at this early stage of the 1000 Genomes Project [102], this pharmacogene database demonstrates the potential impact of these data on pharmacogenomic discovery [42,111]. For example, a novel common SNP located in dihydropyrimidine dehydrogenase (*DPYD*) was found to be associated with hydroxyurea response in the CEU samples [42]. To leverage the available resources, the design of this pharmacogene database was made compatible with the PharmGKB, thereby allowing it to be integrated into the PharmGKB in the future [43,112].

Next wave of pharmacogenomic discovery

The 1000 Genomes Project [102] will greatly expand the sample size and target populations compared with the HapMap Project [7,8,101]. The availability of 11 diverse populations from the HapMap Phase 3 panel, therefore, offers unprecedented opportunities to compare complex phenotypes (e.g., gene expression and drug response) across important current human populations that are relevant to real-world patient demography (e.g., African-Americans and Mexican-Americans in the USA). By contrast, previous studies on gene expression [27] and drug response [28,29] have primarily focused on the three original HapMap populations, which, although supposedly representing a large proportion of world populations, may not reflect the complex genetic structure of certain populations. For example, the ancestry of African-Americans is predominantly from Niger-Kordofanian (~71%), European (~13%) and other African (~8%) populations, although admixture levels vary considerably among individuals [47].

Microarray platforms have proved to be powerful tools for profiling whole-genome gene expression. Taking advantage of the 1000 Genomes Project genotypic data [102], the profiling of gene expression in some HapMap 3 samples could provide novel insights into questions surrounding population differences in gene expression as well as the genetic architecture of gene expression. Although previous studies using the HapMap samples demonstrated some important findings (e.g., common SNPs that account for gene expression variation between populations), only approximately 30% of differentially expressed genes were found to be accounted for by allele frequency differences of either *cis*- or *trans*-acting SNPs [15,48,49].

Although it is possible that other gene regulatory mechanisms, such as miRNA [50] and epigenetics (e.g., DNA methylation) [51], could be responsible for the gene expression variation, no doubt a more detailed map of genetic variation from the 1000 Genomes Project [102] can be used to comprehensively evaluate the contribution of both common and relatively rarer SNPs to gene expression variation. Since gene expression is an intermediate phenotype that sits between DNA sequence variation and higher-level cellular or whole-body phenotypes (e.g., disease susceptibility and individualized drug response), novel insights into gene variation and regulation could enhance our understanding of the observed differences in complex traits including drug response between individual patients and different human populations.

Current pharmacogenomic studies were designed to identify common genetic variants, especially SNPs, which have been found to account for up to 30–50% of the observed variation in drug response [27–29]. Although other mechanisms, such as CNVs and nongenetic factors, could contribute to the remaining fraction of drug response variation [52], a more thorough evaluation of the contribution of genetic variants to drug response could be achieved using the more comprehensive genotypic data from the 1000 Genomes Project [102]. Particularly, given the current sample size and design of the 1000 Genomes Project (e.g., <100 samples for a population) [102], these data may allow identification of certain rare variants with relatively large effects. As the statistical power to identify associations depends on allele frequency, sample size and effect size, the current size may still not have enough power for identifying rare variants only with moderate or minor effects. The current GWAS using the HapMap data largely ignored the contribution of rare SNPs, although their effects on drug response have been appreciated in previous studies [53,54]. The current findings from association studies are likely to be just proxies to the causal genetic variants. Functional validations of these associated SNPs have often been very challenging. A more detailed map of human genetic variation will provide the possibility or promise to locate the true causal variants. Therefore, by the integration of 1000 Genomes Project data and the systematic phenotyping (e.g., mRNA expression profiling, miRNA expression profiling, and drug response phenotyping) of these new samples covered by the project, it will be possible to identify new candidate loci (both previously untapped common and rare ones) and pinpoint causal variants.

Future perspective

In medicine, clinical response to drugs may range widely within and among human populations. Personalized medicine has the benefits of providing safer dosing options, assisting physicians to make better treatment decisions, and facilitating clinical trials in target patient populations. Although it is still in its early stages, pharmacogenomic studies using the HapMap LCL model have demonstrated its potential and promise to identify genetic variants associated with drug cytotoxicity [27–29]. For the next wave of pharmacogenomic discovery and follow-up translational research, however, challenges raised from the current cell-based models must be addressed. The 1000 Genomes Project, aiming to provide a detailed map of genetic variation in over 1000 individuals worldwide, could greatly expand the scope and depth of the current studies by increasing sample size, number of representative populations and the coverage of both common and rare genetic variants [102]. One potential limitation for using these data immediately is that the 1000 Genomes Project is still at its early stage [102]. For example, some scientists question the early-stage data (mostly low coverage at $\sim 2\times$) for how accurate the finished genomes will be (e.g., missing genomic regions and rare variants), given its short timeline and low budget, as well as the lack of phenotypic information (e.g., medical records and basic data such as weight and height) [55]. Although some criticisms (e.g., lack of medical information) are arguable, the 1000 Genomes Project [102] has plans to evaluate the effect of coverage depth and perform regional comparisons with other deep resequencing projects [106] such as the Encyclopedia of DNA Elements (ENCODE) project [56,114]. Therefore,

reasonable expectations are that, after some careful quality control and with progress in the next phase of the 1000 Genomes Project [102] (e.g., ~20× coverage to be used to sequence some protein-coding regions [106]), the final 1000 Genomes Project data will have an acceptable level of genomic coverage as well as high accuracy of allele calls.

Understandably, as a result of the amount of data (estimated at >2 TB) that will be made available from the 1000 Genomes Project [102], efficient bioinformatics tools will be needed to store and accommodate these data to facilitate their use in pharmacogenomic studies. For example, a pharmacogene database enhanced by the early release of the 1000 Genomes Project data offers a convenient way to utilize these data on some well-characterized candidate pharmacology-related genes [42,111]. In addition, previous pharmacogenomic studies have focused on common genetic variants (individual SNPs). To take advantage of the more detailed 1000 Genomes Project data, novel statistical algorithms or data analysis approaches may be necessary to study the contribution of relatively rare variants, CNVs and indel calls from the 1000 Genomes Project to drug response. Pharmacogenomic studies focusing beyond common SNPs will improve our understanding of the genetic basis of drug response. The diverse sample coverage of the 1000 Genomes Project [102,106] can also allow comparisons of genetic factors responsible for drug response between human populations and, therefore, may potentially help identify any race- or population-specific genes and/or variants important for drug response. Furthermore, because of its comprehensiveness, the 1000 Genomes Project data can potentially be used to impute the currently available HapMap genotypic data. For example, only approximately 1 million SNP genotypic data are available for the HapMap Phase 3 samples [101]. These samples, therefore, may be imputed to untyped SNPs using the 1000 Genomes Project data, thus providing a more comprehensive genomic coverage for future studies using these samples, including pharmacogenomic studies. Prospectively, besides gene expression, whole-genome profiling of other molecular targets (e.g., DNA methylation [57] and miRNA expression) on these samples could help build a much more comprehensive drug response model by integrating various ‘-omics’ data. Finally, to leverage the power of existing tools, it is expected that findings based on the 1000 Genomes Project (e.g., a pharmacogene database [42,111]) will be fully integrated into resources such as the PharmGKB and UCSC Genome Browser [58,115]. In summary, the 1000 Genomes Project [102] will be an important resource for the next wave of pharmacogenomic discovery, will greatly enhance our understanding of the genetic basis of drug response, and will prove to be a major step forward on the road of personalized medicine.

Executive summary

- The genetic basis of drug response, a complex trait, may be studied by both candidate-gene approach and genome-wide association approach.
- Current cell-based approaches for pharmacogenomic marker discovery based on their association with drug-induced cytotoxicity have certain limitations, some of which will be addressed with advancement of technologies (e.g., untyped genetic variants).
- Next-generation sequencing technologies will allow large-scale deep resequencing of samples.
- The 1000 Genomes Project aims to provide a much more detailed map of genetic variation on more than 1000 samples from worldwide populations.
- Some bioinformatics tools exist to view and utilize these data, although future bioinformatics development and integration with current resources will be necessary.

- The same samples used in the 1000 Genomes Project have been, and continue to be, used in pharmacogenomic cell-based models such that the unparalleled genetic-variation data released and greater diverse sample coverage will increase the utility of the models.
- Although there are some limitations and concerns (e.g., data accuracy and storage), the 1000 Genomes Project will greatly benefit the pharmacogenomic research community (e.g., imputation of existing samples and comparisons across human populations).

Bibliography

Papers of special note have been highlighted as:

▪ of interest

▪▪ of considerable interest

1. Verstappen CC, Heimans JJ, Hoekman K, Postma TJ. Neurotoxic complications of chemotherapy in patients with cancer: clinical signs and optimal management. *Drugs* 2003;63(15):1549–1563. [PubMed: 12887262]
2. Daugaard G. Cisplatin nephrotoxicity: experimental and clinical studies. *Dan Med Bull* 1990;37(1): 1–12. [PubMed: 2178884]
3. Chatelut E, Delord JP, Canal P. Toxicity patterns of cytotoxic drugs. *Invest New Drugs* 2003;21(2): 141–148. [PubMed: 12889735]
4. Hoskins JM, Carey LA, McLeod HL. *CYP2D6* and tamoxifen: DNA matters in breast cancer. *Nat Rev Cancer* 2009;9(8):576–586. [PubMed: 19629072]
5. Zhou S. Clinical pharmacogenomics of thiopurine *S*-methyltransferase. *Curr Clin Pharmacol* 2006;1 (1):119–128. [PubMed: 18666383]
6. Biason P, Masier S, Toffoli G. *UGT1A1**28 and other *UGT1A* polymorphisms as determinants of irinotecan toxicity. *J Chemother* 2008;20(2):158–165. [PubMed: 18467239]
7. The International HapMap Consortium. The International Hapmap Project. *Nature* 2003;426(6968): 789–796. [PubMed: 14685227]
- 8 ▪▪. The International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437 (7063):1299–1320. Reports the genotypic data from the HapMap Project (Phase 1). [PubMed: 16255080]
9. Van Eerdewegh P, Little RD, Dupuis J, et al. Association of the *ADAM33* gene with asthma and bronchial hyperresponsiveness. *Nature* 2002;418(6896):426–430. [PubMed: 12110844]
10. Binder EB, Holsboer F. Pharmacogenomics and antidepressant drugs. *Ann Med* 2006;38(2):82–94. [PubMed: 16581694]
11. French D, Yang W, Cheng C, et al. Acquired variation outweighs inherited variation in whole genome analysis of methotrexate polyglutamate accumulation in leukemia. *Blood* 2009;113(19):4512–4520. [PubMed: 19066393]
12. Hartford C, Yang W, Cheng C, et al. Genome scan implicates adhesion biological pathways in secondary leukemia. *Leukemia* 2007;21(10):2128–2136. [PubMed: 17673902]
13. Billeci AM, Agnelli G, Caso V. Stroke pharmacogenomics. *Expert Opin Pharmacother* 2009;10(18): 2947–2957. [PubMed: 19925046]
14. Mangravite LM, Wilke RA, Zhang J, Krauss RM. Pharmacogenomics of statin response. *Curr Opin Mol Ther* 2008;10(6):555–561. [PubMed: 19051133]
- 15 ▪. 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(3):631–640. Describes differences in gene expression between HapMap samples from Caucasian residents with northern and western European ancestry from UT, USA and samples from Yoruba people in Ibadan, Nigeria. [PubMed: 18313023]

16. Zhang W, Duan S, Bleibel WK, et al. Identification of common genetic variants that account for transcript isoform variation between human populations. *Hum Genet* 2009;125(1):81–93. [PubMed: 19052777]
17. Frazer KA, Ballinger DG, Cox DR, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449(7164):851–861. Reports the genotypic data from the HapMap Project (Phase 2). [PubMed: 17943122]
18. Peters EJ, Kraja AT, Lin SJ, et al. Association of thymidylate synthase variants with 5-fluorouracil cytotoxicity. *Pharmacogenet Genomics* 2009;19(5):399–401. [PubMed: 19339911]
19. Watters JW, Kraja A, Meucci MA, Province MA, McLeod HL. Genome-wide discovery of loci influencing chemotherapy cytotoxicity. *Proc Natl Acad Sci USA* 2004;101(32):11809–11814. [PubMed: 15282376]
20. 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(23):9758–9763. Study integrates genetic, gene expression and pharmacologic phenotypes. [PubMed: 17537913]
21. 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(3):427–437. [PubMed: 17701890]
22. Huang RS, Duan S, Kistner EO, et al. Genetic variants contributing to daunorubicin-induced cytotoxicity. *Cancer Res* 2008;68(9):3161–3168. [PubMed: 18451141]
23. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from africans. *Mol Cancer Ther* 2008;7(9):3038–3046. [PubMed: 18765826]
24. Hartford CM, Duan S, Delaney SM, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood* 2009;113(10):2145–2153. [PubMed: 19109566]
25. Aksoy P, Zhu MJ, Kalari KR, et al. Cytosolic 5'-nucleotidase III (*NT5C3*): gene sequence variation and functional genomics. *Pharmacogenet Genomics* 2009;19(8):567–576. [PubMed: 19623099]
26. Li L, Fridley BL, Kalari K, et al. Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers. *PLoS One* 2009;4(11):E7765. [PubMed: 19898621]
27. 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(1):15–23. [PubMed: 18392109]
28. Zhang W, Dolan ME. Use of cell lines in the investigation of pharmacogenetic loci. *Curr Pharm Des* 2009;15(32):3782–3795. [PubMed: 19925429]
29. Welsh M, Mangravite L, Medina M, et al. Pharmacogenomic discovery using cell-based models. *Pharmacol Rev* 2009;61(4):413–429. [PubMed: 20038569]
30. O'Donnell PH, Dolan ME. Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy. *Clin Cancer Res* 2009;15(15):4806–4814. [PubMed: 19622575]
31. Zhang W, Dolan ME. On the challenges of the HapMap resource. *Bioinformatics* 2008;24(6):238–239. [PubMed: 18317571]
32. Cheung VG, Conlin LK, Weber TM, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* 2003;33(3):422–425. One of the initial studies to demonstrate natural variation in gene expression in lymphoblastoid cells. [PubMed: 12567189]
33. He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, Xue Y. Geographical affinities of the hapmap samples. *PLoS One* 2009;4(3):E4684. [PubMed: 19259268]
34. Takeuchi F, Serizawa M, Kato N. Hapmap coverage for SNPs in the japanese population. *J Hum Genet* 2008;53(1):96–99. [PubMed: 18043864]
35. Tantoso E, Yang Y, Li KB. How well do hapmap SNPs capture the untyped SNPs? *BMC Genomics* 2006;7:238. [PubMed: 16982009]
36. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;9:387–402. [PubMed: 18576944]
37. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol* 2009;25(4):195–203. [PubMed: 19429539]

38. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;323(5910):133–138. [PubMed: 19023044]
39. Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. *Science* 2008;320(5872):106–109. [PubMed: 18388294]
40. Huang W, Marth G. Eagleview: A genome assembly viewer for next-generation sequencing technologies. *Genome Res* 2008;18(9):1538–1543. [PubMed: 18550804]
41. Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S. Mapview: visualization of short reads alignment on a desktop computer. *Bioinformatics* 2009;25(12):1554–1555. [PubMed: 19369497]
42. •. Gamazon ER, Zhang W, Huang RS, Dolan ME, Cox NJ. A pharmacogene database enhanced by the 1000 genomes project. *Pharmacogenet Genomics* 2009;19(10):829–832. Presents the potential utility of the 1000 Genomes Project data in pharmacogenomics and pharmacogenetics. [PubMed: 19745786]
43. ••. Klein TE, Chang JT, Cho MK, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J* 2001;1(3):167–170. Provides an overview of the Pharmacogenomic Knowledge Base. [PubMed: 11908751]
44. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36(9):949–951. [PubMed: 15286789]
45. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;25(1):25–29. [PubMed: 10802651]
46. •. Gamazon ER, Zhang W, Konkashbaev A, et al. Scan: SNP and copy number annotation. *Bioinformatics* 2010;26(2):259–262. Describes a database with associations at different p-value thresholds between genetic variation and gene expression. [PubMed: 19933162]
47. Tishkoff SA, Reed FA, Friedlaender FR, et al. The genetic structure and history of Africans and African Americans. *Science* 2009;324(5930):1035–1044. [PubMed: 19407144]
48. ••. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet* 2007;39(2):226–231. Demonstrates that specific genetic variation among populations contributes to differences in gene expression phenotypes. [PubMed: 17206142]
49. •. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007;315(5813):848–853. One of the first studies to demonstrate gene expression variation between human populations and the contribution of common genetic variants to gene expression. [PubMed: 17289997]
50. Hobert O. Gene regulation by transcription factors and microRNAs. *Science* 2008;319(5871):1785–1786. [PubMed: 18369135]
51. Wolffe AP, Matzke MA. Epigenetics: regulation through repression. *Science* 1999;286(5439):481–486. [PubMed: 10521337]
52. Ouahchi K, Lindeman N, Lee C. Copy number variants and pharmacogenomics. *Pharmacogenomics* 2006;7(1):25–29. [PubMed: 16354122]
53. Evans WE, Johnson JA. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu Rev Genomics Hum Genet* 2001;2:9–39. [PubMed: 11701642]
54. Freudenberg-Hua Y, Freudenberg J, Winantea J, et al. Systematic investigation of genetic variability in 111 human genes-implications for studying variable drug response. *Pharmacogenomics J* 2005;5(3):183–192. [PubMed: 15809674]
55. Hayden EC. International genome project launched. *Nature* 2008;451(7177):378–379. [PubMed: 18216809]
56. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;306(5696):636–640. [PubMed: 15499007]
57. Zhang W, Huang RS, Dolan ME. Integrating epigenomics into pharmacogenomic studies. *Pharmacogenomics Pers Med* 2008;1(1):7–14.
58. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res* 2002;12(6):996–1006. [PubMed: 12045153]

Websites

101. International HapMap Project. www.hapmap.org
102. The 1000 Genomes Project. www.1000genomes.org
103. NIGMS Human Genetic Cell Repository.
<http://ccr.coriell.org/Sections/Collections/NIGMS/?SsId=8>
104. NIEHS Environmental Genome Project . www.niehs.nih.gov/research/supported/programs/egp/
105. NHGRI Sample Repository for Human Genetic Research .
<http://ccr.coriell.org/Sections/Collections/NHGRI/?SsId=11>
106. 1000 Genomes Project Meeting Report .
www.1000genomes.org/bcms/1000_genomes/Documents/1000Genomes-MeetingReport.pdf
107. NCBI Short Read Archive. www.ncbi.nlm.nih.gov/sites/entrez?db=sra
108. The 1000 Genomes Browser. <http://browser.1000genomes.org>
109. EagleView. <http://bioinformatics.bc.edu/marthlab/eagleview>
110. MapView. <http://evolution.sysu.edu.cn/mapview>
111. A pharmacogene database enhanced by the 1000 Genomes Project.
<http://genemed1.bsd.uchicago.edu/pharmacodb/thougen/main.php>
112. PharmGKB. www.PharmGKB.org
113. SCAN Database. www.scandb.org
114. ENCyclopedia Of DNA Elements. www.genome.gov/10005107
115. UCSC Genome Browser. <http://genome.ucsc.edu>

Table 1

Summary of additional novel SNPs in some candidate pharmacogenes expressed in the samples from Caucasian residents with northern and western European ancestry from UT, USA.

Gene	Full name	Additional novel SNPs [*]	Known SNPs [‡]
<i>ADRB1</i>	Adrenergic, β -1-, receptor	3	2
<i>ADRB2</i>	Adrenergic, β -2-, receptor, surface	0	8
<i>AHR</i>	Aryl hydrocarbon receptor	25	29
<i>ALOX5</i>	Arachidonate 5-lipoxygenase	39	107
<i>BRCA1</i>	Breast cancer 1, early onset	22	65
<i>COMT</i>	Catechol- <i>O</i> -methyltransferase	8	50
<i>CYP2A6</i>	Cytochrome P450, family 2, subfamily A, polypeptide 6	0	1
<i>CYP2D6</i>	Cytochrome P450, family 2, subfamily D, polypeptide 6	1	1
<i>DRD2</i>	Dopamine receptor D2	36	113
<i>GSTP1</i>	Glutathione <i>S</i> -transferase pi 1	2	5
<i>KCNJ11</i>	Potassium inwardly-rectifying channel, subfamily J, member 11	1	7
<i>MTHFR</i>	5,10-methylenetetrahydrofolate reductase	8	54
<i>NQO1</i>	NAD(P)H dehydrogenase, quinone 1	7	16
<i>P2RY1</i>	Purinergic receptor P2Y, G-protein coupled, 1	1	4
<i>SCN5A</i>	Sodium channel, voltage-gated, type V, α -subunit	56	204
<i>SLC19A1</i>	Solute carrier family 19 (folate transporter), member 1	16	38
<i>SULT1A1</i>	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	5	12
<i>TPMT</i>	Thiopurine <i>S</i> -methyltransferase	23	35
<i>TYMS</i>	Thymidylate synthetase	5	42
<i>VDR</i>	Vitamin D receptor	34	111
<i>VKORC1</i>	Vitamin K epoxide reductase complex, subunit 1	1	6

* Number of novel SNPs in addition to those known from dbSNP based on the 35 samples from Caucasian residents with northern and western European ancestry from UT, USA in the 1000 Genomes Project (April 2009).

‡ Number of known SNPs based on dbSNP v129.