Published in final edited form as:

Methods Mol Biol. 2013 ; 1015: 293–310. doi:10.1007/978-1-62703-435-7_19.

A Guide to the Current Web-Based Resources in Pharmacogenomics

Dylan M. Glubb, Steven W. Paugh, Ron H.N. van Schaik, and Federico Innocenti

Abstract

Human genomics research has produced vast amounts of data that can be applied to or used to inform pharmacogenomic studies. The Internet is an extremely useful resource for pharmacogenomics as many Web sites provide access to data from genomic and clinical studies or host tools which can be used to interpret findings or generate hypotheses. Human genetic variation can now easily be explored or visualized through genome browsers and Web-based repositories which store the details of millions of human germ-line and somatic genetic variants. Gene expression data from many different tissue and cell types are available through Web-based repositories, and human genetic variants that associate with mRNA expression can be identified using Web data portals. Pharmacogenetic associations can be explored through publically available data repositories and the functionality of genetic variants predicted through Web-based bioinformatic tools. Furthermore, resources relating to currently used genetic tests are available online. Large clinical and population studies, many linked to medical records, can be queried for the availability of biospecimens or data. In the future, as the amount of genomic and associated clinical data increases, there is little doubt that Web-based resources will continue to evolve and overcome barriers hindering their efficient use, leading to systems-based approaches to pharmacogenomics.

Keywords

Genome browser; Genetic variation; Genotypes; Gene expression; eQTLs; Genetic association studies; GWAS; Biorepositories; Bioinformatic tools

1 Introduction

Since the beginning of the twenty-first century, we have transitioned to a time of routine clinical pharmacogenetic testing with the possibility of routine whole genome sequencing in the near future [1]. Currently, other methods such as DNA oligonucleotide genotyping arrays allow more cost-effective but less comprehensive genome-wide interrogation. Nonetheless, the interpretation and synthesis of genomic data for broad audiences, including patients, physicians and researchers, remains a large hurdle as, even at this point in time, there exists overwhelming amounts of human genomic and pharmacogenomic data.

Web-based resources allow investigators, clinicians and patients to access an extremely wide range of valuable DNA sequence, genotype, phenotype and bioinformatic data that are relevant to pharmacogenomic studies. Searches of PubMed show that these resources are rapidly growing in number (Fig. 1) but the integration of pharmacogenomic-related information remains an issue. One publicly available resource that helps overcome this barrier is the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB; Chapter 20) PharmGKB has changed the way pharmacogenomics research is performed and interpreted, and provide a tremendous resource.

Given the large number of Web-based resources which have relevance to pharmacogenomics, this chapter will only highlight the most useful and comprehensive resources available. It should be noted that in such a quickly changing field, what is described here represents only a snapshot in time as human genomic and pharmacogenomic projects and technologies evolve.

2 Repositories of Germ-line Genetic DNA Sequences and Variants (Table 1)

There are many Web-based genome browsers which enable viewing of reference germ-line DNA sequences of humans and other species and are often integrated with a variety of contextual information. The most well-known include the Ensembl, UCSC and NCBI browsers. These browsers provide annotated reference DNA sequence data and are searchable by chromosomal location, gene or genetic variant, allowing investigators to visualize DNA sequences of interest. The level of annotation varies significantly between the genome browsers with the most comprehensive being the UCSC browser. This browser has many annotations available for display which include those for genetic variation, clinical phenotypes, gene expression and regulation, epigenetics, and comparative genomics.

While genome browsers can provide DNA sequences with genetic variants annotated, they are not themselves repositories of germ-line genetic variation. The largest collection of SNPs can be found in the NCBI dbSNP repository which now contains over 30 million human SNPs. SNPs can be searched for by rs ID or Human Genome Variation Society name and dbSNP provides DNA sequence flanking the SNP of interest, and, when available, population allele and genotype frequencies. dbSNP does not contain information about larger genomic alterations but the Database of Genomic Variants (DGV) fills this gap in the knowledge. DGV lists germ-line insertion/deletion (indel) variants larger than 100 base pairs as well as inversions and copy number variants (CNVs) larger than 1 kilobase. These structural variants can be browsed by chromosome and searches can be performed by DNA sequence, gene or chromosomal location. The Web site also has a genome browser page allowing annotations of sequence with structural variation, SNPs, genes and disease phenotypes. The DGV repository is limited to CNVs observed in healthy humans but dbVar, the NCBI database of genomic structural variation, holds information about germ-line and somatic CNVs identified in healthy and clinical samples. dbVar can be queried by many different criteria including chromosomal position, gene, CNV ID, associated clinical phenotypes, sample type, study and variant size and detailed CNV information is provided.

With regards to genetic variants that are specifically related to pharmacogenetics, the Frequency of Inherited Disorders database (FINDbase) stores allelic frequency data of nearly 150 pharmacogenetic variants from ~ 87,000 individuals belonging to different populations and ethnic groups. FINDbase is described in detail in Chapter 21.

Genome browsers and databases of genetic variation provide information about DNA sequence and variation but these repositories do not contain germ-line DNA genotypes. This role is filled by repositories such as those of HapMap and 1000 Genomes Project. These projects have proved to have many applications to pharmacogenomic studies and have greatly advanced research in the field [2]. HapMap has led the way as a source of human genotype information and has identified millions of common (>5 % minor allele frequency) SNPs and short indels in eleven different ethnic populations. Researchers can use the HaploView software application from the Broad Institute to analyze linkage disequilibrium (LD) patterns using genotype data downloaded from the HapMap Web site. Another way to view LD from HapMap is using the CandiSNPer Web-tool. CandiSNPer can retrieve SNPs of a given HapMap population in a user-defined region flanking the SNP of interest and determine LD between this SNP and plot annotated SNPs according to chromosomal

position. Rarer alleles are not well represented in HapMap but the 1000 Genomes Project aims to describe the map of SNPs, indels and larger structural variants present in at least 0.1 % (for coding regions) to 1 % (for noncoding regions) by whole-genome sequencing of more than 2,000 individuals from five major ethnic populations [3]. 1000 Genomes Project data can be visualized through a specific version of the Ensembl genome browser. Another way to view 1000 Genomes Project data is through the SNP Annotation and Proxy Search (SNAP) Web site. SNAP uses data from 1000 Genomes Project Pilot 1, in addition to HapMap, to allow SNP searches by rs ID and the identification of SNPs based on LD and the generation of LD plots.

There also exist Web-based repositories of smaller targeted studies of human genetic variation such as the NIEHS SNPs and SeattleSNPs projects. The NIEHS SNPs project has two phases: Phase I entailed the identification of SNPs in genes from key biological pathways involved in environmental response through DNA resequencing using a set of human DNA samples representing the diversity of the USA [4]; and Phase II, which is in progress, is using second-generation DNA sequencing to characterize the genetic variation of the entire exome. Variation data are uploaded to the NIEHS SNPs Exome Variant Server which is searchable by gene name or chromosomal location and complete exome sequencing data are available for download. The aim of the SeattleSNPs project is to characterize genetic variation in candidate genes of inflammatory responses by genotyping individuals from HapMap population groups and generates data which is incorporated into the HapMap project. The SeattleSNPs Web site has a Genome Variation Server allowing searches by chromosomal location, gene or rs ID. This database contains 4.5 million variants with corresponding genotype data.

There are several Web-based repositories of germ-line genetic variation from clinical studies. The NCBI database of Genotypes and Phenotypes (dbGaP) contains genotypes, pedigree information, fine mapping results and resequencing traces from over 2,000 clinical datasets. Searches of dbGaP can be made by disease, geno-typing platform or study name or studies can be browsed. The core study of the Wellcome Trust Case Control Consortium (WTCCC) has genotype information derived from genotyping of 500,000 SNPs in 2,000 individuals from each of seven common diseases in addition to 3,000 control individuals from the UK [5]. WTCCC genotype data are not publically available but access can be obtained by researchers after application to the WTCCC Data Access Committee.

Many databases and tools can be used to mine the data originating from these projects, some of which are described extensively in this book. SPSmart is a Web-tool which incorporates genotype data from many of the database repositories such as HapMap, 1000 Genomes and others and allows a specific population-based analysis across databases by SNP, chromosomal location or gene.

3 Repositories of Cancer Somatic Variation (Table 2)

The identification of somatic DNA variation associated with cancer and wide-scale genotyping of such variants has become an area of considerable research activity in recent years. The International Cancer Genome Consortium (ICGC) is collecting genomic data from 50 different cancer types using more than 25,000 individual tumors, initially through whole exome sequencing [6]. The data will include full range of somatic variation (SNPs, indels, CNVs and chromosomal rearrangements) which occur at a frequency of at least 3 % and matching non-tumor tissue will be used to distinguish somatic variants from germ-line ones. The ICGC Data Portal allows searches to be made by genes, samples, simple somatic variants and structural rearrangements. Queries by gene will generate mutation summaries identifying the frequencies of somatic mutations in ICGC tumor datasets. Alternatively,

genes or gene pathways affected by somatic mutations can be identified by searching the database by tumor type. Raw sequencing data can be accessed by researchers upon request. A variety of DNA sequencing and genotyping techniques, which have changed over the course of the project, have been used to identify somatic mutations by The Cancer Genome Atlas (TCGA). TCGA is cataloguing genomic information from more than 20 different types of cancer using matched tumor and normal tissue and plan to collect samples from 500 patients for each cancer. As of early 2012, nearly 6,000 patient samples have been collected and analyzed. The TCGA Data Portal allows searches to be made by genes, somatic variants and disease type, and raw sequencing data can be retrieved and downloaded upon granting of access.

The Catalogue of Somatic Mutations in Cancer (COSMIC) has collated data generated from the aforementioned ICGC and TCGA studies, the Cancer Genome Project (CGP) and targeted sequencing of the NCI60 cell lines (a panel of 60 diverse human cell lines) in known cancer genes, in addition to information extracted through the literature. As of early 2012, this repository holds genotypes from over 600,000 tumors, whole-genome sequencing data from nearly 500 cancer genomes and lists more than 200,000 somatic mutations. The somatic variation data can be searched by gene, sample, tissue or mutation or browsed by tissue and sub-tissue categories to identify genes in which mutations reside and their distribution within a gene. To determine whether somatic mutations in a specific gene may be driving a cancer, the cBio Cancer Genomics Portal is a useful resource. This Web-tool provides the useful function of identifying the percentage of cases within a particular tumor type which display mutations in a gene of interest with data derived from TCGA and the Memorial Sloan-Kettering Cancer Center. A tab linking to survival curve data shows whether there is a significant association of survival and mutations in the gene of interest.

Somatic CNVs are a common feature of cancers and, in addition to COSMIC and TCGA, the Oncomine and Tumorscape repositories contain such information. The Tumorscape Web site allows the querying of data by tumor type or by gene from a study of somatic CNVs in more than 3,000 specimens from 26 types of tumors generated and data can be downloaded from specific studies. Searches by gene will provide summary information of genetic amplifications and deletions in specific cancers while searches by tumor type will summarize the amplifications and deletions identified. The UCSC Cancer Genomics Browser also hosts somatic CNV data in addition to other somatic mutations from the TCGA and other cancer genomic projects. The data is searchable by chromosome or gene but genomic data cannot be currently downloaded.

4 Repositories of Gene Expression Data (Table 3)

Just as technologies have evolved to allow the interrogation of DNA variation at the genome-wide level, it is also now possible to quantitate the mRNA output of the genome through DNA expression arrays or next-generation sequencing (RNA-seq) [7]. One of the biggest resources of these data is the Gene Expression Omnibus (GEO) repository which is maintained on the NCBI Web site. GEO allows the downloading of gene expression data from over 2,500 datasets containing more than 625,000 samples, including many pharmacogenomic studies and can be queried for studies, experimental keywords, genes or even nucleotide sequences. Alternatively, GEO can be browsed by dataset, array platform or samples. Another large repository of mRNA data is ArrayExpress which contains microarray gene expression collected to MIAME standards from over 25,000 experiments. ArrayExpress data can be downloaded upon querying or browsing by studies, experiments, array platforms, or genes. Within the ArrayExpress archive, a subset of more than 5,000 experiments have been curated and re-annotated allowing queries of individual genes to

determine effects on expression across experiments in specific diseases, tissues, cells under different biological conditions or treatments.

Oncomine has a focus on cancer gene expression and includes relevant data from GEO, TCGA and other publically available repositories. Tools in Oncomine allow the comparison of gene expression in tumor and normal tissue samples to identify genes which are specifically expressed in the dataset of interest. The differential analysis option, available in the premium edition of Oncomine, is especially pertinent to the study of pharmacogenomics as it can be used to examine drug sensitivity and patient treatment response data. The Gene Expression across Normal and Tumor tissue (GENT) database also contains gene expression data from tumor and normal tissue. GENT stores mRNA data profiled using Affymetrix microarray gene expression arrays from over 34,000 tissue samples and nearly 1,000 human cancer cell lines. Queries can be made by Affymetrix probe or gene and the output plots gene expression across tissues (tumor and/or normal) or cell lines.

Another source of expression data from normal and tumor tissue samples, including expression of miRNAs, is the UCSC genome browser. Furthermore, gene expression data from the NCI60 cell lines is also available in the browser. The NCI60 cell lines have been used to screen >100,000 chemicals and Cell Miner database contains drug concentrations which inhibit cell growth by 50 % (GI₅₀), expression (gene, protein and miRNA), DNA methylation and fingerprinting data generated from these cell lines. Queries can be made by drug, gene, protein, miRNA, tissue and the corresponding data is available for download. The Systems Biology and Medicine Database (SBM DB) is another repository which stores expression data from cell lines. SBM DB contains mRNA expression profiles from 112 different (normal and tumor) tissues and cell lines measured using Affymetrix U133 microarrays. Searches can be made by gene and relative expression visualized across the different tissues and cells.

5 Repositories of Expression Quantitative Trait Loci (eQTL) Data (Table 4)

Gene expression can be used as a quantitative trait to which SNPs can be associated and such variants are known as expression quantitative trait loci (eQTL). This enables mRNA expression data from genome-wide studies to be analyzed for associations with the large numbers of SNPs which are known today. The field of eQTL research is in the early stages of development but several searchable repositories are available on the Internet and one of them, SCAN, which contains eQTLs identified in lymphoblastoid cell lines (LCLs), is discussed in detail in Chapter 14. The eQTL.uchicago.edu Web site hosts a genome browser which has eQTL annotations from liver, brain, LCL, monocytes, T-cell, and brain studies. The Genotype-Tissue Expression (GTEx) eQTL browser contains eQTL data from liver, brain and LCL studies. Further eQTL data from other tissues will be added as the project progresses. The data can be searched by SNP or gene and results can be filtered by the significance of the eQTL association. eQTLs are also annotated with significant clinical associations observed in Genome-Wide Association Studies (GWAS). seeQTL is a human eQTL browser with data from monocyte, brain and LCL studies and incorporates a meta-analysis approach to score eQTLs across the studies of LCLs from different HapMap populations [8]. The browser is searchable by chromosomal position, gene or SNP.

6 Repositories of Genetic and Genome-Wide Association Studies and Related Tools (Table 5)

Repositories of phenotype–genotype association studies are very useful resources for pharmacogenomic researchers looking to find genes or genetic variants associated with drug responses or clinical relevance for genes or variants of interest. dbGaP is a repository of

such clinical data and can be browsed by studies, clinical variables, analyses and datasets. Genotype and phenotype data are available for download although some data are embargoed or have controlled access. Additionally, dbGaP contains data from GWAS. These analyses allow the testing of associations between clinical phenotypes and a large number of common SNPs. The National Human Genome Research Institute (NHGRI) maintains a catalog which summarizes the findings of GWAS which have assayed at least 100,000 SNPs. The entire catalog can be downloaded or searched by clinical phenotype, chromosomal region, gene or SNP. There are also Web-based portals such as the Phenotype-Genotype Integrator (PheGenI) which allow dbGaP and the NHGRI catalog to be queried by clinical phenotype, SNP, gene or chromosomal location. Queries using these tools will return a summary of significant SNP-phenotype associations with links to the relevant studies in dbGaP or the NHGRI catalog.

The human genome epidemiology (HuGE) Navigator enables the mining of the literature, GWAS, meta-analyses, cancer studies for clinical and epidemiological genetic associations. Queries can be made by disease, gene, SNP, study keyword and the individual databases of associations organized by GWAS, genes and diseases can be downloaded. Searching by disease will generate a table of genes which studies have shown to be associated with that disease and the table further divides studies into GWAS and meta-analyses. There is a bioinformatics tool called Gene Prospector which rates the gene-disease associations based on the available evidence. Links to relevant studies in dbGaP are also provided. The GWAS results link to summaries of the associations. Similarly, searching by gene will generate a table of diseases which studies have shown to be associated with that gene. GWAS Central performs a similar role to the HuGE Navigator and is one of the largest repositories of genetic association studies with summaries of data from over 1,000 studies, including GWAS and candidate gene or region studies. GWAS Central builds upon SNPs and other variants from public to provide findings from genetic association studies without providing individual level genotypes or phenotypes. Summary level data can be presented for one or more studies, comprised of a single or multiple experiments and subjects. Data can be queried or browsed by SNP, gene, chromosomal location, disease, or study keyword and filtered by the p-value of the association.

7 Bioinformatic Tools to Predict SNP Function (Table 6)

Bioinformatic prediction of the functional consequences of a particular genetic variant plays an increasingly important role in genetic association studies, as genetic associations identified from non-candidate driven studies, such as GWAS, do not typically have an obvious mechanistic explanation. Non-synonymous (i.e., amino acid changing) SNPs are the variants most likely to have effects on protein function and many validated pharmacogenetic SNPs are non-synonymous. There are several Web sites which provide a bioinformatic analysis of non-synonymous SNPs. SIFT (sorts intolerant from tolerant) is an algorithm which is based on the degree of amino acid conservation in similar sequences and SNPs can be queried by rs ID or chromosomal location. The results of the analysis are a score and a prediction of whether the SNP is damaging or tolerated by the protein. The Polymorphism Phenotyping (PolyPhen) Web site performs a similar role and contains analyses of the effects of all 54,373 unique human SNPs from the dbSNP build 126 which map to proteins. A score is given to each SNP and a prediction of its effect on parameters such as protein structure, hydrophobicity, and function. The complete database can be downloaded or searches can be made by rs ID or amino acid sequence.

Synonymous SNPs or those not in coding regions can have functional effects, most likely on gene expression, through mechanisms which modify exon–intron splicing, transcription factor and miRNA binding. A useful way to analyze an SNP when uncertain of its

functionality is to use one of the Web-based meta-analysis tools. SNP Function Prediction (FuncPred) from the NIEHS can be queried by SNP, gene, or chromosomal position and provides predictions of transcription factor and miRNA binding, splicing, non-synonymous SNPs and PolyPhen analysis, stop codons and scores for regulatory potential, and sequence conservation based on comparative genomics are generated. In addition, SNPs in LD with the variant of interest in a specific population can also be visualized. F-SNP provides similar functional analyses of effects on splicing and transcription, and also protein function, including posttranslational effects, using information from 16 tools and databases. FastSNP provides a gene or SNP based search and provides a risk definition based upon a decision tree hierarchy classifying SNPs into different categories of functional significance from nonsense to downstream SNPs with no known function.

The Encyclopedia Of DNA Elements (ENCODE) project has generated much genomic and epigenomic data which aids the identification of functional elements in the genome and thus enables hypotheses to be generated about genetic variants. ENCODE data has been generated from experiments examining transcription, chromatin accessibility, histone modification, DNA methylation, transcription factor binding, and several other genomic/epigenomic features in many different cell types [9]. Data are available for download or can be visualized in the UCSC genome browser.

8 Web sites Related to Pharmacogenetic Testing (Table 7)

There are now over a 100 FDA-approved drugs which have pharmacogenetic information in their label and the FDA Web site contains a list of these drugs with links to label and pharmacogenetic information. Warfarin is one of these drugs and its pharmacogenetics has been very well characterized. Indeed, a dosing algorithm for warfarin is available at the Warfarin Dosing Web site (Chapter 22). The algorithm takes into account the genotypes of variants in *CYP2C9*, *CYP4F2*, *GGCX* and *VKORC1*, in addition to clinical factors, to estimate the warfarin dose a patient should receive.

There are more pharmacogenetic tests available than just those listed on FDA drug labels and the Genetic Diagnostic Network (GENDIA), an international network of more than 100 laboratories, currently lists 201 tests on its pharmacogenetic Web site (PHARMACO-GENDIA). A searchable repository of genetic test information is provided by the Genetic Testing Registry (GTR) from the NIH. The GTR links genes and genetic variants to diseases and drug responses. The database can be queried by test, drug, disease and other clinical phenotypes, gene, protein or laboratory providing genetic tests. In the context of cancer genetic tests, Cancer GEM KB maintains a list of related tests which can be queried by disease, gene or drug.

9 Biorepositories and Population-Based Cohorts with Linked Medical Records (Table 8)

Large populations which are accurately clinically phenotyped are increasingly needed to explore and understand the relationship between genetic variants and drug responses. Moreover, studies linked to national biorepositories which store biological samples from participants, such as the BioBank Japan Project [10] and UK Biobank [11], have great potential for pharmacogenomic analysis. The UK Biobank is a longitudinal project which has collected blood, urine and saliva samples, in addition to health and lifestyle information, from 500,000 individuals aged 40–69 who live in the UK. To access samples, researchers have to register, propose a research project and pay charges related to administration of the project. However, researchers can visualize data summaries using UK Biobank Showcase, an open access Web-portal, which allows the study to be queried or browsed for factors such

as lifestyle, family and medical history, health and physical measures. Indeed, many data coordinating centers and registries allow their registries to be searched for numbers of patients with specific clinical characteristics without restrictions. The Cancer Genetics Network (CGN) is a data coordinating center for 14 clinical centers and the database contains information from over 26,000 individuals with cancer and/or a family history of cancer. The CGN database can be queried by demographics, family history of cancer, clinical characteristics, and whether a genetic test has been performed to determine the number of patients who meet the criteria selected. Access to the database is then possible after acceptance of a research proposal. One Web-based repository which contains medical and personal information, in addition to genotype data, with unrestricted access is the Personal Genome Project (PGP) [12]. The PGP has enrolled 1,000 individuals so far and aims to incrementally expand to 100,000.

There are also large studies which have matching electronic medical records. For example, the Electronic Medical Records and Genomics (eMERGE) Network is an NHGRI consortium of five institutions that link DNA from over 100,000 individuals to electronic medical records [13]. DNA samples and electronic medical records from (eMERGE) Network of DNA repositories can be accessed after a successful affiliate membership application that contains a research proposal.

The International Serious Adverse Event Consortium (iSAEC) provides a Web-based portal to serious adverse event clinical and genotype data from participating pharmaceutical companies and academic institutions available to researchers who sign the consortium's data use agreement. In addition, iSAEC has also compiled data from similarly genotyped population controls matched for age, sex and ethnicity.

Web-based repositories which enable the location and availability of biospecimens with associated clinical data provide useful tools for researchers. The NCI Specimen Resource Locator can be queried for tumor specimens and other matching biological samples (e.g., DNA/RNA) based on the type of tumor, the associated clinical data available and will link to specific biospecimen repositories matching the specified criteria. Similarly, the Rare Disease-HUB from the Office of Rare Diseases Research (NIH) can be queried by disease, biospecimen anatomic source, processing and storage method, imaging data available and will then link to repositories matching the specified criteria. The Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) contains information about clinical studies supported by the National Heart Lung and Blood Institute (NHLBI). The BioLINCC can be queried by study conditions and biospecimen type. Researchers must register to request data and biospecimens or add study data and information to the database.

10 Conclusion and Future Directions

Pharmacogenomic research is a rapidly evolving research area moving away from singular in vitro/ex vivo/in vivo experiments and converging into large scale agnostic in vivo genotype/phenotype studies of large cohorts of patients. This growth and evolution is certainly tied to the growth in processing and communication capacity provided by Internet connected workstations and computer facilities which dwarf the technology available only a few years ago. Barriers remain to the utilization and interpretation of the results of these experiments including the lack of a common standardized application programming interface among pharmacogenomic repositories which leads to a lack of connectivity and some level of redundancy. Problems of ethics, human subject protection, and the commercialization of publicly funded resources also remain important areas for discussion among both scientific and public groups. As new technologies for massive data generation further emerge and become more widely used the expected population of computational

Web-based portals for genomic data analysis will also grow and perhaps led to the emergence of a new type of system science to merge these data sources together into knowledge.

References

1. Wagner MJ. Pharmacogenetics and personal genomes. *Per Med*. 2009; 6:643–652. [PubMed: 20190862]
2. Zhang W, Dolan ME. Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. *Pharmacogenomics*. 2010; 11:249–256. [PubMed: 20136363]
3. Consortium, G.P. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
4. Rieder MJ, et al. The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies. *Drug Metab Rev*. 2008; 40:241–261. [PubMed: 18464045]
5. Consortium, T.W.T.C.C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–678. [PubMed: 17554300]
6. Hudson TJ, et al. International network of cancer genome projects. *Nature*. 2010; 464:993–998. [PubMed: 20393554]
7. Roy NC, et al. A comparison of analog and Next-Generation transcriptomic tools for mammalian studies. *Brief Funct Genomics*. 2011; 10:135–150. [PubMed: 21389008]
8. Xia K, et al. seeQTL: a searchable database for human eQTLs. *Bioinformatics*. 2012; 28:451–452. [PubMed: 22171328]
9. Myers RM, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011; 9:e1001046. [PubMed: 21526222]
10. Nakamura Y. The BioBank Japan project. *Clin Adv Hematol Oncol*. 2007; 5:696–697. [PubMed: 17982410]
11. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. *Pharmacogenomics*. 2005; 6:639–646.
12. Angrist M. Eyes wide open: the personal genome project, citizen science and veracity in informed consent. *Per Med*. 2009; 6:691–699. [PubMed: 22328898]
13. Kho AN, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med*. 2011; 3:79re1.

Fig. 1. Publications related to genomic resources. PubMed searches for associated terms such as “genome browser” (a) and “expression database” (b) show their rapid growth since the beginning of the twenty-first century

Table 1
Repositories of germ-line DNA sequence and genetic variation

Resource	Web address	Description
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/index.html	Serves as a central repository for both single base nucleotide substitutions and short indel polymorphisms
International HapMap Project	http://www.hapmap.org/index.html.en	Repository of SNP genotypes from eleven different ethnic populations
NIEHS SNPs	http://egp.gs.washington.edu	A SNP discovery resource focused on examining the relationships between environmental exposures, inter-individual sequence variation in human genes and disease risk in US populations
SeattleSNPs	http://pga.mbt.washington.edu	A SNP discovery resource focused on genes involved in pathways that underlie inflammatory responses in humans
UCSC Genome Browser	http://genome.ucsc.edu	Repository of annotated genomic reference DNA sequences
Ensembl	http://ensembl.org	Repository of annotated genomic reference DNA sequences
NCBI Genome	http://www.ncbi.nlm.nih.gov/sites/genome	Repository of annotated genomic reference DNA sequences
Wellcome Trust Case Control Consortium	http://www.wtccc.org.uk/info/access_to_data_samples.shtml	Identifies genome sequence variants influencing major causes of human morbidity and mortality, through implementation and analysis of large-scale genome wide association studies
1000 Genomes	http://www.1000genomes.org	Repository of DNA sequencing and variant calls from different ethnic populations
SPSmart	http://spsmart.cesga.es/	Allows browsing and combination of genotypes from many large-scale genomic databases into user defined groups
dbGaP	http://www.ncbi.nlm.nih.gov/gap	Repository of human genotypes from human genotype-phenotype studies
SNAP	http://www.broadinstitute.org/mpg/snap/index.php	Web-based service for retrieval of SNP LD data from the International HapMap and 1000 Genomes Projects
Database of Genomic Variants	http://projects.tcag.ca/variation/	A repository of structural genetic variation found in healthy individuals
Innate Immunity Programs for Genomic Applications	https://regepi.bwh.harvard.edu/IIPGA2/index_html	Repository of human genotypes and haplotypes from genes related to innate immunity
JSNP database	http://snp.ims.u-tokyo.ac.jp/	Repository of common SNPs in the Japanese population
dbVar	http://www.ncbi.nlm.nih.gov/dbvar/	An NCBI repository of CNVs
FINDbase	http://www.findbase.org	Repository of allele frequencies of pharmacogenetic markers in different populations

Table 2
Repositories of somatic (cancer) variation

Resource	Web address	Description
ICGC	http://www.icgc.org/	Repository planned to contain DNA sequencing and somatic mutation data from 50 different tumor types
COSMIC	http://www.sanger.ac.uk/cosmic/	Repository of somatic mutation, genotype and whole genome sequencing data from cancer studies
TCGA	http://cancergenome.nih.gov/dataportal	Repository planned to contain DNA sequencing and somatic mutation data from 20 different tumor types
variant GPS	http://variantgps.nci.nih.gov/cgfseq/pages/home.do	Repository of genotyping data and genetic variants identified from targeted next-generation sequencing in cancer studies
cBio Cancer Genomics Portal	http://www.cbioportal.org/public-portal/	Searchable Web-tool which integrates tumor and somatic mutation data from TCGA and the Memorial Sloan-Kettering Cancer Center
Tumorscape	http://www.broadinstitute.org/tumorscape	Repository of somatic CNV data from multiple cancer types
Oncomine	https://www.oncomine.org	Searchable Web-tool which integrates somatic CNV data from TCGA
Cancer Genomics Browser	https://genome-cancer.ucsc.edu/	Web-tool which integrates somatic mutation data from TCGA and other cancer genomic studies

Table 3
Repositories of gene expression data

Resource	Web address	Description
ICGC	http://www.icgc.org/	Copy number, rearrangement, expression, and mutation data
GEO	http://www.ncbi.nlm.nih.gov/geo/	Repository of gene expression data from >2,500 studies
Oncomine	http://www.oncomine.org	Repository of gene expression data from GEO, TCGA and other projects
Cell Miner	http://discover.nci.nih.gov/cellminer/	Repository of gene expression and GI ₅₀ drug concentration data from NCI60 cell lines
SBM DB	http://www.lsbm.org/site_e/database/	Repository of mRNA expression from healthy and tumoral tissues
ArrayExpress	http://www.ebi.ac.uk/arrayexpress/	Repository of microarray-derived mRNA expression data from >25,000 studies
GENT	http://medical-genome.kribb.re.kr/GENT/	Repository of microarray-derived mRNA expression from >34,000 tissue samples
Cancer Genome Anatomy Project	http://cgap.nci.nih.gov/cgap.html	Repository of gene expression from normal, precancer and cancer cells
UCSC Genome Browser	http://genome.ucsc.edu	DNA sequences annotated with gene expression data from a wide range of sources

Table 4
Repositories of eQTL data

Resource	Web address	Description
SCAN	http://www.scandb.org	Repository of SNP and CNV eQTLs identified in LCLs
eqtl.uchicago.edu	http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/	Genome browser with eQTL annotations
GTEx	http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi	Repository of eQTL data from different tissues

Table 5
Repositories of GWAS and clinical genetic associations

Resource	Web address	Description
GWAS Central	http://www.gwascentral.org/	Repository of genetic association studies
HuGE Navigator	http://hugenavigator.net/HuGENavigator/home.do	Series of Web-tools which enable mining of the literature and genetic association studies
OMIM	http://omim.org/	Compendium of human genes and related phenotypes
NHGRI catalog	http://www.genome.gov/gwastudies/	Repository of GWAS results
dbGaP	http://www.ncbi.nlm.nih.gov/gap	Repository of phenotype–genotype data from clinical studies
PheGenI	http://www.ncbi.nlm.nih.gov/gap/PheGenI	Web-tool which allows queries of dbGaP and the NHGRI catalog

Table 6
Tools to predict SNP function

Resource	Web address	Description
ENCODE Project	http://genome.ucsc.edu/ENCODE/	ENCODE data can be directly downloaded or visualized through the UCSC Genome Browser
SNP Function Portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx	Provides annotation of SNPs at the genome, transcript, protein, pathway, disease and population levels
FastSNP	http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp	Web-tool which incorporates bioinformatic analyses of SNP function
pfSNP	http://pfs.nus.edu.sg	Web-tool which incorporates bioinformatic analyses of SNP function and results from genetic association studies
Pupasuite	http://pupasuite.bioinfo.cipf.es	Web-tool which incorporates bioinformatic analyses of SNP function
SIFT	http://sift.jcvi.org/	Web-tool which predicts effects of non-synonymous SNPs
PolyPhen	http://genetics.bwh.harvard.edu/pph/	Web-tool which predicts effects of non-synonymous SNPs
FuncPred	http://snpinfo.niehs.nih.gov/snpfunc.htm	Web-tool which incorporates bioinformatic analyses of SNP function
F-SNP	http://compbio.cs.queensu.ca/F-SNP/	Web-tool which incorporates bioinformatic analyses of SNP function

Table 7
Web sites related to pharmacogenetic testing

Resource	Web address	Description
FDA	http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics	Links to pharmacogenomic information
PHARMACO-GENDIA	http://www.pharmaco-gendia.net	Repository of pharmacogenetic tests
GTR	http://www.ncbi.nlm.nih.gov/gtr/	Repository of genetic test information
CancerGEM KB	http://www.hugenavigator.net/CancerGEMKB	Repository of cancer genetic tests
Warfarin Dosing	http://www.warfarindosing.org/Source/Home.aspx	Web-based algorithm for warfarin dosing
PharmGKB	http://pharmgkb.org	Repository of pharmacogenetic information and PGx-based dosing guidelines

Table 8
Web sites for biorepositories and large cohorts with linked medical information

Resource	Web address	Description
CGN	http://www.cancergen.org/	Clinical database of patients from 14 cancer centers
NCI Specimen Resource Locator	http://pluto3.nci.nih.gov/tissue/default.cfm	Web- portal which allows searches for biological samples from cancer patients
BioLINCC	http://biolincc.nhlbi.nih.gov/home/	Web-portal which allows searches for biological samples and clinical data from NHLBI studies
Rare Disease-HUB	http://biospecimens.ordr.info.nih.gov/	Web-portal which allows searches for biological samples from patients with rare diseases
PGP	http://www.personalgenomes.org/	Open access repository of genotype data with linked medical and personal information
iSAEC	http://www.saeconsortium.org	Repository of clinical and genotyping data from studies of adverse drug events
eMERGE Network	https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page	Network of DNA repositories with linked electronic medical records
UK Biobank	http://www.ukbiobank.ac.uk/	Population-based repository of biological samples with linked medical, lifestyle and family history data