BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 2 2010, pages 259–262
doi:10.1093/bioinformatics/btp644

Databases and ontologies

SCAN: SNP and copy number annotation

Eric R. Gamazon[1], Wei Zhang[1], Anuar Konkashbaev[1], Shiwei Duan[1], Emily O. Kistner[2], Dan L. Nicolae[1,3], M. Eileen Dolan[1] and Nancy J. Cox[1,4,*]

[1]Department of Medicine, [2]Department of Health Studies, [3]Department of Statistics and [4]Department of Human Genetics, The University of Chicago, Chicago, IL, USA

Received on September 11, 2009; revised on November 2, 2009; accepted on November 12, 2009

Advance Access publication November 17, 2009

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Genome-wide association studies (GWAS) generate relationships between hundreds of thousands of single nucleotide polymorphisms (SNPs) and complex phenotypes. The contribution of the traditionally overlooked copy number variations (CNVs) to complex traits is also being actively studied. To facilitate the interpretation of the data and the designing of follow-up experimental validations, we have developed a database that enables the sensible prioritization of these variants by combining several approaches, involving not only publicly available physical and functional annotations but also multilocus linkage disequilibrium (LD) annotations as well as annotations of expression quantitative trait loci (eQTLs).

Results: For each SNP, the SCAN database provides: (i) summary information from eQTL mapping of HapMap SNPs to gene expression (evaluated by the Affymetrix exon array) in the full set of HapMap CEU (Caucasians from UT, USA) and YRI (Yoruba people from Ibadan, Nigeria) samples; (ii) LD information, in the case of a HapMap SNP, including what genes have variation in strong LD (pairwise or multilocus LD) with the variant and how well the SNP is covered by different high-throughput platforms; (iii) summary information available from public databases (e.g. physical and functional annotations); and (iv) summary information from other GWAS. For each gene, SCAN provides annotations on: (i) eQTLs for the gene (both local and distant SNPs) and (ii) the coverage of all variants in the HapMap at that gene on each high-throughput platform. For each genomic region, SCAN provides annotations on: (i) physical and functional annotations of all SNPs, genes and known CNVs within the region and (ii) all genes regulated by the eQTLs within the region.

Availability: http://www.scandb.org

Contact: ncox@medicine.bsd.uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Association studies of complex diseases and pharmacogenomic studies, along with recent advances in our ability to survey hundreds of thousands of single nucleotide polymorphisms (SNPs) on high-throughput genotyping platforms, highlight the need for characterizing and prioritizing a list of polymorphisms potentially implicated in disease susceptibility or therapeutic drug response. The International HapMap project (The International HapMap Consortium, 2003) (http://www.hapmap.org/) was launched as an international effort to catalog common genetic variants in human populations. The HapMap Project has released genotypic information on >3.1 millions SNPs of 270 Epstein–Barr Virus transformed lymphoblastoid cell lines (LCLs) (Frazer *et al*., 2007) derived from apparently healthy individuals of African, Asian and European ancestry. This important development reflects the ever-increasing amount of genotype and, uniquely to HapMap, haplotype information available in the public domain on polymorphisms in the human genomes.

Together with the availability of these cell lines, the HapMap resource has proven to be of tremendous value in assisting researchers to identify genetic determinants responsible for complex traits or phenotypes (Zhang *et al*., 2008a). Notably, this resource makes it possible to do unsupervised genome-wide studies to assess the contribution of common genetic variants including SNPs and copy number variants (CNVs) to gene expression. Gene expression is itself a complex trait, but also acts as an intermediate phenotype between the genetic loci and higher level cellular or clinical phenotypes, such as disease risk or individual drug response. Particularly, variation in gene expression level within a single population (Morley *et al*., 2004; Stranger *et al*., 2005) or between populations (Spielman *et al*., 2007; Stranger *et al*., 2007; Zhang *et al*., 2008b) has been mapped to the human genome as expression quantitative trait loci (eQTLs), suggesting that common genetic variants including SNPs and CNVs contribute to a substantial fraction of the natural variation in gene expression.

Using gene expression phenotype data (mRNA level) generated from the Affymetrix GeneChip® Human Exon 1.0 ST Array, we performed family-based QTDT (Quantitative Transmission Disequilibrium Test) analysis (Abecasis *et al*., 2000a; Abecasis *et al*., 2000b) on over 13 000 transcript clusters (gene level) with reliable expression—that is, the $\log_2$-transformed expression signal is >6 in at least 80% of the samples—and over 2 million common SNPs with minor allele frequency (MAF) >5% and no Mendelian inheritance transmission errors in the set of HapMap trios of CEU (Caucasians of northern and western European ancestry from UT, USA) and YRI (Yoruba people from Ibadan, Nigeria) samples, evaluated separately (Duan *et al*., 2008a). Each transcript cluster includes a set of

*To whom correspondence should be addressed.

probesets (exon level) containing all known exons and 5′- and 3′-untranslated regions (UTRs) in the genome. Since SNPs in probes can result in false-positive eQTL signals (Alberts *et al*., 2007), SNP data from dbSNP (Sherry *et al*., 2001) (build 129) were used to identify probes that hybridize to regions containing SNPs; such probes were excluded from the expression analyses (Duan *et al*., 2008b). SCAN uses summary analyses of HapMap SNP associations to transcriptional expression to annotate polymorphisms.

Haplotype data from HapMap is also revealing the structure of linkage disequilibrium (LD) in the human genome. Recent work in multilocus LD (Nicolae, 2006) provides a framework for interpreting results from genome-wide association studies (GWAS) by quantifying, for any set of markers (e.g. SNPs), the coverage of each of the high-throughput platforms relative to a reference panel (e.g. HapMap SNPs). Even with the advances in genotyping technologies, it is likely that the causative loci are not genotyped (Zhang and Dolan, 2008). There is thus a need to integrate untyped variants into testing for association with a complex trait or a pharmacological phenotype. Multilocus LD, which may be calculated with the use of HapMap haplotype frequencies, provides a computationally feasible way to measure how much the typed variants capture the available information (i.e. how little redundancy is present). Such information, for example, can be put to use in the design of assayable SNPs, in the choice of genotyping platform for candidate genes or in the reduction of redundancy.

A distinguishing feature of SCAN in its present implementation is, therefore, the integration of gene expression and (multilocus and pairwise) LD information, not simply physical and functional annotations characteristic of public databases, in characterizing and prioritizing genetic variants.

2 IMPLEMENTATION

The SCAN database has been implemented using a software solution stack known as LAMP. The acronym refers to the use of Linux as operating system, Apache as web server, MySQL as SQL management system and PHP as scripting language. In addition to the web infrastructure developed on LAMP, additional software modules and scripts were written in Perl and C++ to process off-line datasets (Tables 1 and 2) coming from such diverse public domain projects as dbSNP (Sherry *et al*., 2001), RefSeq Database (Pruitt *et al*., 2007), Entrez Gene, Database of Genomic Variants (Iafrate *et al*., 2004) and HapMap (The International HapMap Consortium, 2003) and from such commercial entities as Affymetrix (http://www.affymetrix.com) (Affymetrix, Inc., Santa Clara, CA, USA) and Illumina (http://www.illumina.com) (Illumina, Inc., San Diego, CA, USA), and to generate (multilocus) LD and genotype–gene expression association datasets not available elsewhere.

SCAN is built on an extensible and modular architecture (Supplementary Figure 1). Its use of LAMP, for example, makes it perfectly suited to integrate data from such diverse data sources as Gene Ontology (http://www.geneontology.org/) (Ashburner *et al*., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) (Kanehisa *et al*., 2004). SCAN's flexible database schema makes it possible to incorporate data from other GWAS, not just the expression phenotype associations in LCLs that constitute the initial datasets, to annotate variants.

Table 1. SCAN coverage

	Description
Genes	RefSeq Database, Entrez Gene
Transcript clusters	Affymetrix GeneChip Human Exon 1.0 ST Array
HapMap	Release 23a
dbSNP	Build 129
CNVs	Database of Genomic Variants Build 36 (hg18)
High-throughput platforms	Affymetrix, Illumina, Perlegen

Table 2. LD datasets

High-throughput platforms
Affymetrix Genome-Wide Human SNP Array 6.0
Affymetrix Genome-Wide Human SNP Array 5.0
Affymetrix GeneChip Human Mapping 100K
Illumina's High Density Human 1M-Duo
Illumina's Human610-Quad Infinium HD BeadChip
Illumina's HumanHap550-Duo BeadChip

3 METHODS

3.1 SNP Query

At present, our SNP Query supports RefSNP (Pruitt *et al.*, 2007) (identified by 'rs' numbers) and Affymetrix SNP identifiers as required input. Since subsequent analysis on the result set may be performed, we provide query results in a variety of useful formats: HTML, comma-delimited (.csv) or tab-delimited text file. A SNP Query can be refined to use optional parameters to define the returned annotation (Supplementary Figure 2A):

(1) General SNP information such as SNP genomic coordinates (chromosome and base position) mapped to the reference assembly and the SNP's RefSeq alleles.

(2) Host gene (using RefSeq genomic coordinates for the gene) as well as SNP function, using dbSNP's classification scheme, that indicates whether a SNP represents a coding change (e.g. a 'nonsense' causing changes to a STOP codon, a frameshift indel change or an amino acid substitution), intronic, a 5′ or 3′ splice site, a 5′- or 3′-UTR, within 5′ 2 kb to a gene or within 3′ 0.5 kb to a gene.

(3) Left- and right-flanking genes.

(4) The genes showing local and distant associations to the SNP in the CEU, YRI and combined samples, along with the *P*-values calculated by the QTDT (Abecasis *et al.*, 2000a; Abecasis *et al.*, 2000b).

Each SNP (in dbSNP) is clickable and returns a screen that displays three types of information: general, population specific and LD (Supplementary Figure 2B). The general information includes the SNP's 'ss' identifier in dbSNP, base position, host gene, function, possible RefSeq mRNA and protein products, ancestral allele and dbSNP validation methods (Sherry *et al.*, 2001) (i.e. how the variant is ascertained through a non-computational method). dbSNP updates are handled in SCAN via dbSNP's RsMergeArch table, which contains dbSNP's 'rs merge' history. RefSeq mRNA and protein products are themselves clickable, and through the NCBI Entrez Programming Utilities (a URI-based application programming interface) yields additional real-time NCBI information. Population-specific information currently includes MAFs in the different populations, which

were calculated using the HapMap bulk data (Hapmap release 23a). The data in the LD section of the screen, available for HapMap SNPs, are generated with TUNA (Nicolae, 2006) and show how well the variant is interrogated on the high-throughput genotyping platforms using multilocus LD coefficients, the maximum pairwise LD coefficient r2 between each SNP in HapMap and the typed SNPs within 200 kb of the HapMap variant, whether the SNP is on the platform and the typed SNPs used in the imputation.

3.2 Gene Query

Our Gene Query supports (case-insensitive) official Entrez gene symbols as required input. In the future, we will add RefSeq Gene IDs (Pruitt *et al.*, 2007) as supported input. Different output formats are provided, as in the SNP Query, for possible use in downstream analysis. Gene filtering criteria specified by the user include (Supplementary Figure 3A):

(1) The gene's genomic coordinates mapped to the reference assembly.

(2) dbSNP variants within and up to a user-specified length (in kilobases) from the gene.

(3) The list of eQTLs that predict the expression of the gene at a user-defined *P*-value threshold in user-specified population and MAF.

(4) The option to display only eQTLs on the same chromosome.

Each gene in the result set is clickable and returns a screen that displays general information on the gene as well as coverage information on each of the high-throughput genotyping platforms (Supplementary Figure 3B). The general information section of the Gene screen provides official Entrez Gene ID, gene type (e.g. protein coding or microRNA), a description, genomic coordinates relative to the reference assembly, orientation, map location, status (e.g. validated, reviewed or predicted) and other frequently used designations for the gene. The gene's structure—the coding regions, UTRs, intronic regions and the existence of alternative splicing—is graphically shown as are the gene's position and strand orientation relative to neighboring genes using the NCBI's Entrez URI-based graphical application programming interfaces (APIs).

3.3 Multilocus LD

To generate the multilocus LD datasets, we downloaded Affymetrix, Illumina and Perlegen annotation data files for each of the platforms. At present, we have calculated multilocus LDs for the following platforms: Affymetrix Genome-Wide Human SNP Array 6.0, Affymetrix Genome-Wide Human SNP Array 5.0, Affymetrix GeneChip Human Mapping 100K, Illumina's High Density Human 1M-Duo, Illumina 650K, Illumina 550K and Perlegen 330. We applied TUNA (Nicolae, 2006) to these datasets using HapMap population panel data in the CEU and YRI populations. For each HapMap SNP, we provide multilocus LD coefficients as well as the typed SNPs on certain high-throughput platforms used to calculate the multilocus LD. We calculated different measures for a given gene by taking the average, median and the multilocus LD at Q1 and at Q3 of the multilocus LD of all the HapMap SNPs within and up to 2 kb of the gene. This approach allows us to study the multilocus LD distribution across every gene.

4 DISCUSSION

SCAN currently supports queries from three primary interfaces:

(1) A SNP Query that retrieves physical and functional annotations, host and flanking genes, and the genes whose expressions are predicted to be regulated, at a user-specified *P*-value threshold, by the variant in the CEU, YRI and the combined CEU and YRI samples.

(2) A Gene Query that obtains all dbSNP (Sherry *et al.*, 2001) variants (build 129) within and up to a user-specified distance (in kilobases) of the gene, maps the gene to its genomic coordinates relative to the reference assembly and returns the list of local (*cis*-) and distant (*trans*-acting) regulators of the gene. The SNPs located within the 4 mb of a gene were defined as local SNPs; other SNPs (including those on other chromosomes) were defined as distant SNPs (Duan *et al.*, 2008a).

(3) A Genomic Region Query that returns the list of dbSNP (Sherry *et al.*, 2001) variants in the specified genomic region (NCBI build 36), the list of all genes located with the region and all genes whose expressions are regulated by the SNPs within the region at a user-specified *P*-value threshold in the CEU, YRI and the combined CEU and YRI samples.

Each of the primary interfaces allows batch upload of SNP, gene or genomic region lists.

We are developing an application programming interface (API) in order to support the building of genomics applications that utilize SCAN and to enable the bioinformatics community to integrate SCAN into existing tools. The technical specification, written in Simple Object Access Protocol, facilitates the exchange of structured information with other databases or with client applications, and can be used in conjunction with other web protocols such as HTTP. This programmatic approach saves application developers from building a gene expression/LD calculation engine that is available through SCAN and enables the integration of SCAN's datasets in real time. Indeed, to facilitate pharmacogenomic studies, we are collaborating with PharmGKB (http://www.pharmgkb.org) to integrate SCAN's datasets into PharmGKB's interface seamlessly, using the described API.

In summary, the SCAN database allows user-friendly queries of the results of GWAS on the association of HapMap variants with gene expression at user-specified thresholds. SCAN also uses multilocus measures of disequilibrium to summarize some of the reported LD relationships among SNPs and to characterize coverage of genes by high-throughput genotyping platforms. SCAN annotates SNPs not only with physical and functional information currently distributed across several public databases but also with extent of LD and the ability to predict transcript expression. The current version of SCAN is built upon the genotypic and phenotypic data generated on the HapMap LCLs, which have some intrinsic limitations (e.g. only one tissue type, limited sample size, cell line collection time biases, low coverage of rarer SNPs). Interpretation of results based on SCAN may require taking into account these factors. Expanding SCAN using data on other tissues (currently in development) and from some ongoing research efforts such as the 1000 Genomes Project, as well as integrating other gene regulation mechanisms such as DNA methylation (Zhang *et al.*, 2008c) and microRNA may provide a more comprehensive database in the future.

ACKNOWLEDGEMENTS

The authors are grateful to members from the Dolan Lab and from the Cox Lab for testing the database and providing helpful feedback.

Funding: Pharmacogenetics of Anticancer Agents Research Group (http://pharmacogenetics.org) (grant U01GM61393) from the National Institute of General Medicine; University of Chicago Breast Cancer Spore (P50 CA125183) funded by the National

Cancer Institute; ENDGAMe (ENhancing Development of Genome-wide Association Methods) initiative (U01 HL084715); The University Of Chicago DRTC (Diabetes Research and Training Center) (P60 DK20595).

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al.* (2000a) A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.*, **66**, 279–292.

Abecasis,G.R. *et al.* (2000b) Pedigree tests of transmission disequilibrium. *Eur. J. Hum. Genet.*, **8**, 545–551.

Alberts,R. *et al.* (2007) Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE*, **2**, e622.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Duan,S. *et al.* (2008a) Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.*, **82**, 1101–1113.

Duan,S. *et al.* (2008b) SNPinProbe_1.0: a database for filtering out probes in the Affymetrix GeneChip(R) Human Exon 1.0 ST array potentially affected by SNPs. *Bioinformation*, **2**, 469–470.

Frazer,K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Kanehisa,M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Morley,M. *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Nicolae,D.L. (2006) Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet. Epidemiol.*, **30**, 718–727.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Spielman,R.S. *et al.* (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.*, **39**, 226–231.

Stranger,B.E. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.

Stranger,B.E. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Zhang,W. and Dolan,M.E. (2008) Beyond the HapMap genotypic data: prospects of deep resequencing projects. *Curr. Bioinform.*, **3**, 178–182.

Zhang,W. *et al.* (2008a) The HapMap Resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform. Biol. Insights*, **2**, 15–23.

Zhang,W. *et al.* (2008b) Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.*, **82**, 631–640.

Zhang,W. *et al.* (2008c) Integrating epigenomics into pharmacogenomic studies. *Pharmacogenomics Pers. Med.*, **1**, 7–14.