BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2540–2542
doi:10.1093/bioinformatics/btu240

Databases and ontologies

Advance Access publication May 7, 2014

GEPdb: a database for investigating the ternary association of genotype, gene expression and phenotype

Daeun Ryu[1,†], SeongBeom Cho[2,†], Hun Kim[1], Sanghyuk Lee[1,3] and Wankyu Kim[1,3,*]

[1]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, [2]Division of Biomedical Informatics, National Institute of Health, Korea Center for Disease Control, Cheongwon 363-951 and [3]Ewha Global Top5 Research Program, Ewha Womans University, Seoul 120-750, Republic of Korea

Associate Editor: John Hancock

ABSTRACT

Summary: GEPdb integrates both genome-wide association studies and expression quantitative trait loci information, the two primary sources of genome-wide mapping for genotype–phenotype and genotype–expression associations together with phenotype-associated gene lists. The GEPdb provides simultaneous interpretation of both genetic risks and potential gene regulatory pathways toward phenotypic outcome by establishing the ternary relationship of genotype–expression–phenotype (GEP). The analytic scope is further extended by linkage disequilibrium from five different populations of the international HapMap Project.

Availability and implementation: http://ercsbweb.ewha.ac.kr/gepdb.

Contact: wkim@ewha.ac.kr

Received on January 21, 2014; revised on April 16, 2014; accepted on April 18, 2014

1 INTRODUCTION

Genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) are the principal methods for genome-scale mapping of genotype–phenotype and genotype–expression relationships, respectively. In most cases, GWAS provides little information on the underlying mechanism at the molecular and cellular levels. A recent survey showed that 5∼16% of the GWAS hits fall into coding regions (Schaub *et al.*, 2012) allowing direct interpretation of protein structural and functional impacts. The majority of hit single nucleotide polymorphisms (SNPs) may have some regulatory function through transcriptional regulation, miRNA targeting or epigenetic changes. Therefore, if there are common SNPs shared by both GWAS and eQTL, the eQTL genes are good candidates for the underlying pathological process and its phenotypic outcome (Cookson *et al.*, 2009).

Many dedicated databases have been developed for either GWAS (Hindorff *et al.*, 2009; Li *et al.*, 2012; Thorisson *et al.*, 2009) or eQTL (Lonsdale *et al.*, 2013; Xia *et al.*, 2012; Yang *et al.*, 2010). There are few integrated databases for both GWAS and eQTL data that (i) support a mechanistic interpretation of GWAS results, and (ii) infer causal genetic variations responsible for differential expression of disease signature genes via eQTL mapping. We developed *GEPdb* that seamlessly integrates a comprehensive collection of GWAS, eQTL and phenotype-associated gene lists. Such integration provides a novel opportunity to analyze the ternary relationship of genotype–expression–phenotype (GEP) in a genome scale. The hit SNPs (tagging SNPs) may belong to a haplotype block, whereas their neighboring SNPs may be actually functional in both GWAS and eQTL mapping by linkage disequilibrium (LD). GEPdb supports expansion of SNPs using LD information.

2 DATA COLLECTION AND PROCESSING

We compiled and processed datasets for (i) >47 million SNPs, (ii) >1300 GWAS, (iii) ∼0.8 million eQTLs and (iv) >17 million eQTL SNP-gene pairs, as well as ∼12 000 phenotype-associated gene signatures categorized by traits, disease subtypes and tissue/cell types. The summary of the datasets is shown in Table 1, and the detailed statistics are shown online at the GEPdb Web site. The SNPs are categorized as *gene*, *5′UTR*, *exon*, *intron*, *3′UTR* and *intergenic*. Notably, all GWAS datasets and the phenotype-associated gene lists are classified by *Trait*, *MeSH*, *Study*, *FirstAuthor*, *Disease Class* and/or *tissue* by a substantial manual annotation and are organized into a hierarchical tree-like structure.

3 SYSTEM OVERVIEW

GEPdb integrates a comprehensive collection of GWAS (genotype–phenotype), eQTL (genotype–expression) and phenotype-associated gene signatures. The schematic overview is shown in Figure 1. Using the GEPdb, a list of SNPs or genes can be interpreted by associating the ternary relationships of GEP. The input SNPs or genes are selected among the precompiled datasets of GEPdb or can be directly entered by the user. The analytical procedure is designed to be simple, linear and bidirectional, i.e. G:E, G>E>P or P>E>G mode (Fig. 1). In the G:E mode, a list of SNPs and a gene set are entered as input, and then the SNPs and genes are linked to each other via the eQTL relationship. Alternatively, the user can select gene signature(s) for a particular trait (or disease), and check whether they are linked to the GWAS hit SNPs for the same or related phenotype via eQTL mapping. Any SNP set can be expanded using the LD information among the five different populations of the International Hapmap Project phases I–III. In the G>E>P mode, a list of

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. Data statistics compiled in the GEPdb

Data	Source	Number of unique SNPs (Genes)	Number of datasets
SNP	dbSNP Build 137	47 491 458	–
	dbNSFP v2.0b4	753 614	–
GWAS	GWAS catalog (GC)	9966	741 (1338)[b]
	HuGe Navigator (HN)[a]	7266	642 (1137)
QTL	seeQTL	262 632	1 158 424[c]
	eQTL Browser	101 115	120 921
	GTEx	20 987	26 694
	Ding (Ding *et al.*, 2010)	14 414	15 726
	Dixon (Dixon *et al.*, 2007)	387 877	15 695 524
Gene signature	GeneSigDB	15 990	1131[b]
	GAD	14 721	2671
	FunDO	4053	411
LD information	ASW	871 238	3 240 154[d]
	CEU	2 400 299	42 393 828
	CHB	2 256 741	42 754 104
	JPT	2 220 673	43 350 226
	YRI	2 236 433	18 155 106

[a]Redundant entries with GWAS catalog are removed. [b]Number of phenotypes (GWAS studies); [c]Number of SNP-gene pairs; [d]Number of SNP pairs by LD (R^2>0.8).

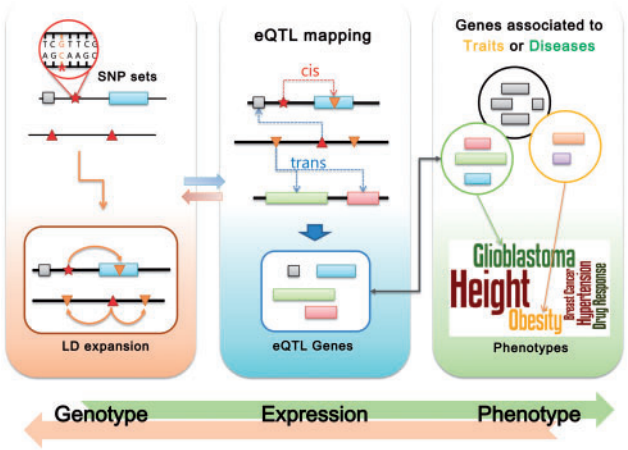

Fig. 1. The analytic flow of information in GEPdb

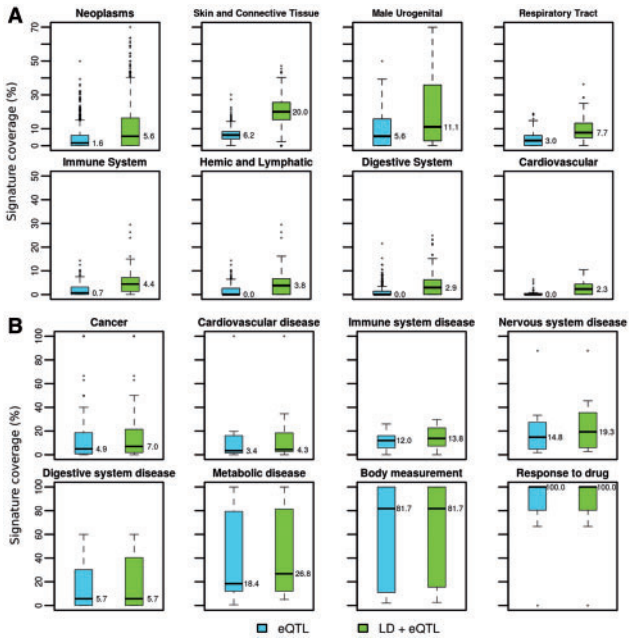

Fig. 2. The coverage of gene signatures from (**A**) GeneSigDB and (**B**) GAD by the GWAS hit SNPs via eQTL mapping (blue) and LD + eQTL mapping (green). The GWAS hit SNPs are merged together for each disease if there are multiple GWAS for a single disease

input SNPs are entered and linked to the downstream regulated genes via eQTL (G>E). Then, these genes are compared against the precompiled gene sets related to various diseases and other phenotypes (E>P). The summary result is displayed as table, in which the statistical significance of the overlap between gene or SNP sets (*P*-value and *Q*-value) is calculated by the hypergeometric distribution and the Bonferroni correction method. The direction of analysis is reversed in the P>E>G mode. Starting from a list of genes [e.g. typically, differentially expressed genes (DEGs) for a disease: P>E] as input, their associated eQTL SNPs are retrieved as potentially regulatory loci.

4 DISCUSSION

A critical question regarding the GEP ternary relationship is to what extent the expression signatures (DEGs) can be interpreted by GWAS hit SNPs, via (i) eQTL and further by (ii) LD information. Among the datasets compiled in the GEPdb, we analyzed paired sets of the GWAS hit SNPs and the gene signatures from GAD and geneSigDB annotated by the same disease using MeSH terms and manual inspection. The coverage of the signature genes by GWAS hit SNPs via eQTL (+LD) varied highly depending

on the disease category and individual study. The median coverage was 0∼100 and 2.3∼100% by eQTL and LD + eQTL, respectively, and some signature genes resulted in no outcome (Fig. 2). The overall low coverage may, at least partially, result from inaccurate mapping between GWAS and the corresponding expression signatures because of disease subtypes or heterogeneity. Nevertheless, a significant fraction of known gene signatures were traced back to their potential genetic risks using GEPdb (G:E mode). Similarly, any set of gene signatures or candidate SNPs can also be analyzed using the G>E>P or P>E>G modes to generate testable hypotheses.

Funding: Research Program funded by the Korea Centers for Disease Control and Prevention (2013-E71001-00:4800-4851-307), the National Research Foundation of Korea (NRF-2011-0014992, NRF-2013M3A9B6046519, NRF-2012M3A9C5048707) and GIST Systems Biology Infrastructure Establishment Grant through ERCSB.

Conflict of Interest: none declared.

REFERENCES

Cookson,W. *et al.* (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.

Ding,J. *et al.* (2010) Gene expression in skin and lymphoblastoid cells: refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.*, **87**, 779–789.

Dixon,A. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Li,M.J. *et al.* (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Schaub,M.A. *et al.* (2012) Linking disease associations with regulatory information in the human genome. *Genome Res.*, **22**, 1748–1759.

Thorisson,G.A. *et al.* (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.

Xia,K. *et al.* (2012) seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, 451–452.

Yang,T.-P. *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.