BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 3 2012, pages 451–452
doi:10.1093/bioinformatics/btr678

Databases and ontologies

Advance Access publication December 13, 2011

seeQTL: a searchable database for human eQTLs

Kai Xia[1,*], Andrey A. Shabalin[1], Shunping Huang[2], Vered Madar[1], Yi-Hui Zhou[1], Wei Wang[2], Fei Zou[1], Wei Sun[1,3], Patrick F. Sullivan[3] and Fred A. Wright[1,*]

[1]Department of Biostatistics, [2]Department of Computer Science and [3]Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: seeQTL is a comprehensive and versatile eQTL database, including various eQTL studies and a meta-analysis of HapMap eQTL information. The database presents eQTL association results in a convenient browser, using both segmented local-association plots and genome-wide Manhattan plots.

Availability and implementation: seeQTL is freely available for non-commercial use at http://www.bios.unc.edu/research/genomic_software/seeQTL/.

Contact: fred_wright@unc.edu; kxia@bios.unc.edu

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

Received on September 21, 2011; revised on November 2, 2011; accepted on November 27, 2011

1 INTRODUCTION

The association of RNA expression traits with DNA variation, including from single-nucleotide polymorphisms (SNPs) and copy-number variants (CNVs), has been a subject of active inquiry in recent years, shedding light on fundamental biological processes underlying transcription. Here we use the generic term 'expression quantitative trait loci' (eQTLs) to describe these DNA variants and their associated expression traits (Feuk *et al.*, 2006). A large number of studies have been published on HapMap lymphoblastoid cell lines and other human tissues, covering several continental-level populations (Choy *et al.*, 2008; Dimas *et al.*, 2009; Grundberg *et al.*, 2009; Montgomery *et al.*, 2010; Myers *et al.*, 2007; Pickrell *et al.*, 2010; Price *et al.*, 2008; Schadt *et al.*, 2008; Spielman *et al.*, 2007; Stranger *et al.*, 2007; Zeller *et al.*, 2010). In addition to the general importance of dissecting transcriptional regulation, eQTL analysis may also provide a window into the mechanisms underlying transcription-mediated disease (Consoli *et al.*, 2002).

Several online databases are available which report eQTL associations based on published datasets. SCAN (Gamazon *et al.*, 2009) is a large-scale database of genetic and genomic data, which allows users to search for eQTLs by querying multiple genes or SNPs/CNVs, but is not designed primarily as an eQTL database. The eQTL Browser (Pickrell *et al.*, 2010) is based on the Gbrowse platform (Donlin, 2009) and displays results from multiple studies and allows navigation throughout the genome. The GTEx (Genotype-Tissue Expression) eQTL database will be populated by tissue-specific eQTL information as the GTEx project (http://www.ncbi.nlm.nih.gov/gtex/test/GTEX2/) progresses, but is currently limited in navigability. SNPexp (Holm *et al.*, 2010) provides the database for users to investigate a specified region, but contains a limited number of eQTL datasets. Despite the current attention to eQTL datasets, the need remains for a powerful and versatile eQTL database to easily investigate regions, loci and transcripts of interest.

Here we introduce *seeQTL*, a new database of human eQTL associations. It is based on the Gbrowse2 platform, which is more powerful and customizable than the original Gbrowse. Most of the studies represented in seeQTL (Supplementary Material) were re-analyzed using our own pipeline, combining quality control, population stratification control, association testing and false discovery rate (FDR) control (Fig. 1) (Benjamini and Hochberg, 1995). In addition, we performed a meta-analysis to obtain a consensus association score for each eQTL across the HapMap studies and populations. Here we use the terms 'cis eQTL' for local eQTLs (within 1 Mb of a gene) and 'trans eQTL' for more distant eQTLs. *Cis* associations are displayed using either *segment* plots (Fig. 2) or FDR q-value association Manhattan plots and *trans* associations using Manhattan plots.

1.1 Datasets and analysis

We collected 14 human eQTL datasets, including unrelated HapMap lymphoblastoid cell lines (Choy *et al.*, 2008; Dimas *et al.*, 2009; Montgomery *et al.*, 2010; Pickrell *et al.*, 2010; Price *et al.*, 2008; Spielman *et al.*, 2007; Stranger *et al.*, 2007), human cortical samples (Myers *et al.*, 2007) and monocytes (Zeller *et al.*, 2010). The gene expression data were downloaded from NCBI GEO, and genotype data were downloaded from HapMap or the authors' website (Supplementary Material). We excluded a sample for low expression quality and excluded SNPs with low minor allele frequency (MAF). Detail of eQTL calculations and FDR control are provided in the Supplementary Material and Figure 1. Summarized results of the datasets are provided in Supplementary Table S1. Additional datasets will be added as data are made available. We soon anticipate loading results from the 'godot' study, an eQTL evaluation of peripheral blood gene expression in ~800 monozygotic and 750 dizygotic twin pairs.

1.2 Consensus method

The HapMap lymphoblastoid cell line data consist of multiple expression datasets and cover several continental-level populations. Separate analyses can be performed within each dataset. However, as the data are all from the same tissue source, the availability

*To whom correspondence should be addressed.

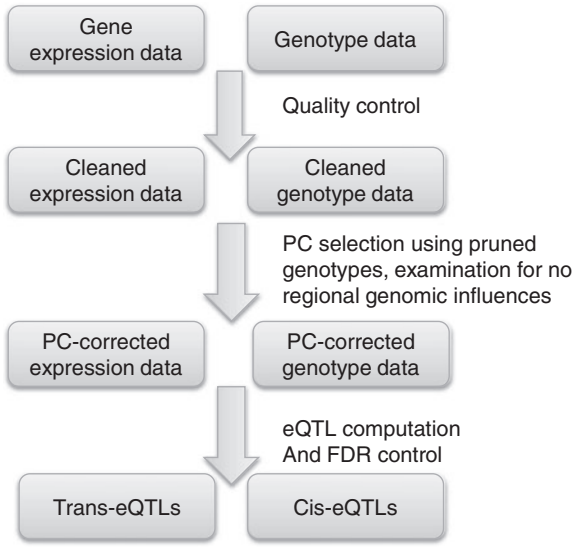

Fig. 1. The pipeline of eQTL analysis. PC, principal component.

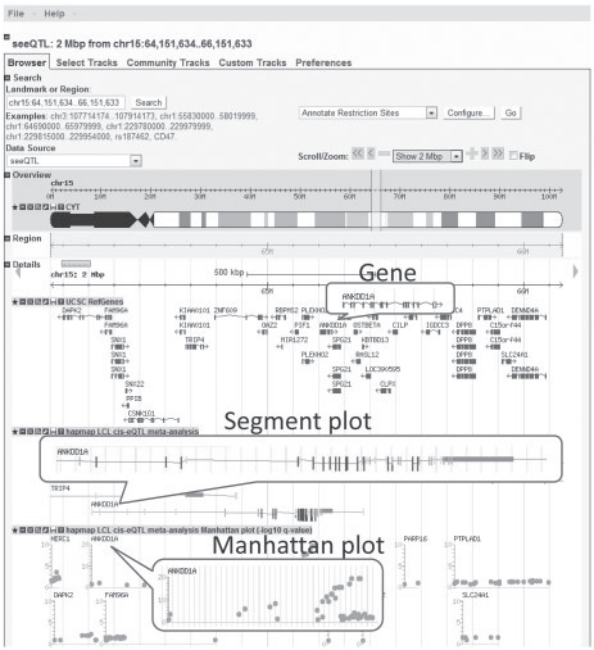

Fig. 2. A screenshot of the seeQTL browser, with various features highlighted: eQTLs can be viewed as segment plots for *cis*-associations, showing SNPs as 'connected' to associated genes, or as Manhattan plots for both *cis* and *trans* associations.

of a single consensus meta-analysis would greatly facilitate eQTL analysis of HapMap samples. We applied a standard meta-analysis approach to obtain a consensus score for each transcript and each SNP with study-specific weights chosen to maximize power (Supplementary Material).

1.3 Usage features

The seeQTL browser is navigable using text-searches for genes and SNPs, presenting a table view of features containing these text strings. Alternatively and subsequently, seeQTL is navigable by clicking and zooming. These browser features allow maximum flexibility in focusing on specific genes and genomic regions. As described above, *cis*-associations are displayed using segment plots, which are useful to display the 'connection' between genes and associated SNPs, as well as Manhattan plots. For the SNPs in a region, all the genes to which these SNPs exhibit significant association can also be displayed in Manhattan plots. Tracks based on individual datasets are shown separately, as well as the consensus HapMap track. Comparisons of seeQTL features and advantages are shown in Supplementary Table S2.

ACKNOWLEDGEMENTS

We thank Guanhua Chen for advice and help with the database development.

Funding: Gillings Innovation Laboratory in Statistical Genomics (NIMH 5RC2MH089951 and R01MH090936).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser*, **B 57**, 289–300.

Choy,E. *et al.* (2008) Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, **4**, e1000287.

Consoli,L. *et al.* (2002) QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Mol. Biol.*, **48**, 575–581.

Dimas,A.S. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.

Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9 9.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Gamazon,E.R. *et al.* (2009) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.

Grundberg,E. *et al.* (2009) Population genomics in a disease targeted primary cell model. *Genome Res.*, **19**, 1942–1952.

Holm,K. *et al.* (2010) SNPexp - A web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels. *BMC Bioinformatics*, **11**, 600.

Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

Myers,A.J. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.

Pickrell,J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Price,A.L. *et al.* (2008) Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genet.*, **4**, e1000294.

Schadt,E.E. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.

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.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Zeller,T. *et al.* (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS One*, **5**, e10693.