PanCanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types

Jing Gong1,2, Shufang Mei1, Chunjie Liu2,3, Yu Xiang2, Youqiong Ye2, Zhao Zhang2, Jing Feng2, Renyan Liu4, Lixia Diao5, An-Yuan Guo3, Xiaoping Miao1,* and Leng Han2, *

1Department of Epidemiology and Biostatistics, Key Laboratory of Environmental Health of Ministry of Education, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, PR China, 2Department of Biochemistry and Molecular Biology, McGovern Medical School at The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, 3Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, PR China, 4Department of Pharmacology, State University of New York Upstate Medical University, Syracuse, NY 13210, USA and 5Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

ABSTRACT

Expression quantitative trait locus (eQTL) analysis, which links variations in gene expression to genotypes, is essential to understanding gene regulation and to interpreting disease-associated loci. Currently identified eQTLs are mainly in samples of blood and other normal tissues. However, no database comprehensively provides eQTLs in large number of cancer samples. Using the genotype and expression data of 9196 tumor samples in 33 cancer types from The Cancer Genome Atlas (TCGA), we identified 5 606 570 eQTL-gene pairs in the cis-eQTL analysis and 231 210 trans-eQTL-gene pairs in the trans-eQTL analysis. We further performed survival analysis and identified 22 212 eQTLs associated with patient overall survival. Furthermore, we linked the eQTLs to genome-wide association studies (GWAS) data and identified 337 131 eQTLs that overlap with existing GWAS loci. We developed PancanQTL, a user-friendly database (http://bioinfo.life.hust.edu.cn/PancanQTL/), to store cis-eQTLs, trans-eQTLs, survival-associated eQTLs and GWAS-related eQTLs to enable searching, browsing and downloading. PancanQTL could help the research community understand the effects of inherited variants in tumorigenesis and development.

INTRODUCTION

Single nucleotide polymorphisms (SNPs), the most common type of human genetic variation, play important roles in human complex traits and diseases (1–3). Genome-wide association studies (GWAS) identified more than 10 000 SNPs associated with susceptibility of human traits or diseases (4,5). Most GWAS-detected risk SNPs are located in the genome’s non-coding regions (6), indicating that these SNPs mainly exert their functional roles via regulating gene expression. Therefore, understanding SNP regulation of gene expression is essential for interpreting disease related SNPs.

Expression quantitative trait locus (eQTL) analysis, which links variations in gene expression to genotypes, has been demonstrated as a powerful approach to understanding the effects and molecular mechanism of functional SNPs (7–10). Previous studies identified eQTLs mainly from lymphoblastoid cell lines and normal human tissues (9,11–13). For example, the Genotype-Tissue Expression (GTEX) consortium identified eQTLs from 7051 tissue samples of 44 tissues from 449 donors (13). Due to the significance of eQTLs, several databases have been developed to collect eQTLs, including the GTEX Portal (13), ExSNP (14), seeQTL (15) and SCAN (16). However, no database comprehensively provides eQTLs in large number of cancer samples. The majority of eQTLs identified from cancer samples are cancer-specific through a comparison between tumor and normal samples (17). Therefore, it is necessary to analyze eQTLs from large-scale cancer samples to further understand the functional effects of eQTLs in cancer. Furthermore, the majority of studies and databases neglected trans-eQTLs, which are highlighted with significant functions in recent studies (7,18). Collectively, systematic and large-scale investigations of both cis- and trans-eQTLs in multiple cancer types would provide the research community with a further understanding of inherited variant effects in tumorigenesis and development.

*To whom correspondence should be addressed. Tel: +1 713 500 6039; Fax: +1 713 500 0652; Email: leng.han@uth.tmc.edu
Correspondence may also be addressed to Xiaoping Miao. Tel: +86 27 8365 0744; Fax: +86 27 8365 0744; Email: miaoxp@mail.hust.edu.cn
The Cancer Genome Atlas (TCGA) generated a large amount of omics data, including RNA sequencing, genotype data and clinical survival information from more than 10,000 cancer samples. These data provide a valuable source for eQTL analysis and further integrative analysis across different cancer types.

DATA COLLECTION AND PROCESSING

Genotype data collection, imputation and processing

To comprehensively identify eQTLs across different cancer types, we obtained genotype data from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/), which detected the genotypes using Affymetrix SNP 6.0 array containing 898,620 SNPs. To increase the power for eQTL discovery, we imputed autosomal variants for all samples in each cancer type using IMPUTE2 (19), with 1000 Genomes Phase 3 (20) as the reference panel. To improve computation efficiency, we used the two-step procedure of IMPUTE2, which includes pre-phasing, and the imputation of the phased data. After imputation, we used the following criteria to select SNPs (13): (i) imputation confidence score, INFO ≥ 0.4, (ii) minor allele frequency (MAF) ≥ 5%, (iii) SNP missing rate <5% for best-guessed genotypes at posterior probability ≥ 0.9 and (iv) Hardy–Weinberg Equilibrium P-value > 1 × 10^{-6} estimated by Hardy–Weinberg R package (21) (Figure 1A).

Gene expression data collection and processing

The gene expression profiles were obtained from the TCGA data portal (https://gdc-portal.nci.nih.gov/), which contains 20,531 genes for each sample. In each cancer type, genes with average expression (RSEM calculated by Expectation-Maximization (22)) of ≥ 1 were retained. To minimize the effects of outliers on the regression scores, the expression values for each gene across all samples were efficiently minimization of outliers on the regression scores, the expression values for each gene across all samples were transformed into a standard normal based on rank (13) (Figure 1C).

Covariates

Previous studies showed that factors affecting global gene expression may reduce the eQTL-identifying power (23,24). To remove the global effects on gene expression, covariates are usually included in eQTL analyses (9,13). To remove the effect of population structure on gene expression, we used smartpca in the EIGENSOFT program (25) to perform principal component (PC) analyses for each cancer type, and selected the top five PCs in genotype data as covariates. To remove the hidden batch effects and other founders in the expression data, we used PEER software (26) to select the first 15 PEER factors from expression data as covariates. To remove the potential effects of clinical status on gene expression, age (9), gender (13) and tumor stage (17) were included as additional covariates (Figure 1B).

Identification of eQTLs

For each cancer type, the genotype data, expression data and covariates were processed to three N (genotype, expression or covariates) × S (samples) matrix files with matched sample order. The gene location (hg19) was downloaded from Genomic Data Commons (https://gdc.cancer.gov/). The SNP location (hg19) was downloaded from dbSNP (https://www.ncbi.nlm.nih.gov/projects/snp/) (v137). eQTL analysis was performed by Matrix eQTL (27) in linear regression model. SNPs with false discovery rates (FDR) < 0.05 were defined as eQTLs. Cis-eQTLs were defined if the SNP was within 1 Mb from the gene transcriptional start site (TSS) (13), and trans-eQTLs were defined if the SNP was beyond that point (Figure 1D).

Survival-associated eQTLs

Many genes are associated with cancer prognoses (28), and eQTLs may influence the prognosis by altering gene expression. To identify survival-associated eQTLs, we examined the associations between eQTLs and patient overall survival. For each eQTL, samples were classified into three groups: homozygous genotype AA, heterozygous genotype Aa and homozygous genotype aa (A and a represent two alleles of one SNP). The log-rank test was used to examine the differences in survival time, and Kaplan–Meier (KM) curves were plotted to represent the survival time for each group. eQTLs with FDR < 0.05 were defined as survival-associated eQTLs (Figure 1D).

GWAS-related eQTLs

Risk SNPs identified in GWAS studies were downloaded from the GWAS catalog (http://www.ebi.ac.uk/gwas/) (5). GWAS linkage disequilibrium (LD) regions were extracted from SNAP (https://personal.broadinstitute.org/plinn/snap/ldsearch.php) (29) with parameters (SNP dataset: 1000 Genomes; r^2 (the square of the Pearson correlation coefficient of linkage disequilibrium) threshold: 0.5; population panel: CEU (Utah Residents with Northern and Western European Ancestry); Distance limit: 500 kb), eQTLs that overlap with GWAS tagSNPs and LD SNPs (r^2 ≥ 0.5) were identified as GWAS-related eQTLs.

DATABASE CONTENT AND USAGE

Samples in PancanQTL

PancanQTL included 9196 tumor samples from 33 cancer types. The sample size of each cancer type ranged from 36 in cholangiocarcinoma (CHOL) to 1092 in breast invasive carcinoma (BRCA) (Table 1). For the genotype data, we obtained on average 4,480,214 SNPs for each cancer type after imputation and quality control, ranging from 2,765,921 for BRCA to 5,245,402 for acute myeloid leukemia (LAML). After removing lowly expressed genes (RSEM < 1), there were on average 17,814 genes for each cancer type, ranging from 16,758 for uveal melanoma (UVM) to 18,790 for testicular germ cell tumors (TGCT).

eQTLs in PancanQTL

For each cancer type, the average associations of ~81 billion SNP-gene pairs were tested for cis- and trans-eQTL mapping. In cis-eQTL analysis, we identified 5,606,570 eQTL-gene pairs in 33 cancer types at a per-tissue FDR < 0.05,
Cis-eQTLs and trans-eQTLs across 33 cancer types

Genotyping array data of 9196 TCGA samples

Imputed SNPs

Quality control:
1) Imputation confidence score INFO ≥ 0.4
2) Minor allele frequency ≥ 5%
3) SNP missing rate < 5% for best-guessed genotypes at posterior probability ≥ 0.9
4) Hardy-Weinberg equilibrium p > 1×10⁻⁶

Covariates of population structures

EIGENSOFT

mRNA sequencing data of 9196 TCGA samples

Covariates of expression data

Covariates of clinical status

PEER

17 814 genes per cancer

Quality control:
Gene with average expression (RSEM) ≥ 1

17 814 genes per cancer

Preprocessing:
Standard normalization based on rank for each gene

Survival-associated eQTLs

GWAS data from GWAS catalog

Clinical data of TCGA samples

Cis-eQTLs and trans-eQTLs across 33 cancer types

Matrix eQTL

Web design and interface

Results were organized into a set of relational MySQL tables (30), with the website constructed using HTML and PHP. We designed four modules to display cis-eQTLs, trans-eQTLs, survival-associated eQTLs and GWAS-related eQTLs (Figure 2A). Users could browse each eQTL module simply by clicking the corresponding module. On the home page, we designed an advanced search box for a comprehensive query across four modules (Figure 2B). For example, the user can select a cancer type (e.g. STAD) and input an SNP ID (e.g. rs2351010), gene symbol (e.g. ERAP2) or genomic region (e.g. chr1:1–1000000) to search eQTLs in four modules. A quick search option is available on each page (top right) to search by SNP ID, gene symbol or genomic region. Users can download cis-eQTLs and trans-eQTLs for each cancer type from the ‘Download’ page. The ‘Help’ page provides information for data collection and processing. PancanQTL welcomes any feedback by email on the ‘Contact’ page.
Table 1. Summary of eQTLs for each cancer type in PancanQTL.

| Cancer typea | No. of samples | No. of genes | No. of genotypes | Pairs | egenes | eQTLs | Pairs | egenes | eQTLs |
|--------------|----------------|--------------|------------------|-------|-------|-------|-------|-------|-------|
| ACC          | 77             | 17,562       | 3,678,145        | 4610  | 222   | 4,558 | 984   | 60    | 957   |
| BLCA         | 408            | 18,171       | 4,242,910        | 142,562 | 5573 | 120,374 | 9199 | 1575 | 3114 |
| BRCA         | 1092           | 17,991       | 2,765,921        | 438,476 | 11,859 | 317,935 | 73,124 | 6013 | 20,466 |
| CESC         | 300            | 17,975       | 4,367,017        | 95,702 | 4165 | 84,484 | 2209 | 674 | 971 |
| CHOL         | 36             | 17,767       | 4,106,282        | 11 | 2 | 11 | 5011 | 127 | 4436 |
| COAD          | 286            | 17,500       | 4,576,984        | 164,356 | 5048 | 145,461 | 3085 | 373 | 2359 |
| DLBC          | 48             | 17,245       | 4,945,365        | 391 | 15 | 391 | 5 | 3 | 5 |
| ESCA          | 184            | 18,372       | 4,563,674        | 39,358 | 1603 | 36,589 | 425 | 56 | 410 |
| GBM           | 150            | 17,650       | 4,660,522        | 59,788 | 1901 | 55,855 | 481 | 55 | 465 |
| HNSC          | 518            | 17,985       | 4,302,347        | 267,797 | 6502 | 228,069 | 9285 | 1064 | 7389 |
| KICH          | 66             | 17,212       | 3,902,792        | 7264 | 320 | 7038 | 5826 | 157 | 4669 |
| KIRC          | 527            | 17,812       | 4,632,879        | 521,072 | 8739 | 410,720 | 13,973 | 948 | 12,200 |
| KIRP          | 290            | 17,715       | 4,951,414        | 186,310 | 4920 | 164,159 | 2712 | 302 | 2516 |
| LAML          | 123            | 17,099       | 5,245,402        | 70,375 | 1758 | 64,696 | 580 | 38 | 397 |
| LGG           | 515            | 17,563       | 4,688,205        | 578,617 | 9177 | 437,580 | 21,236 | 1804 | 13,084 |
| LIHC          | 369            | 18,816       | 4,218,042        | 151,613 | 5723 | 128,956 | 16,675 | 2230 | 3963 |
| LUAD          | 514            | 18,190       | 4,435,432        | 259,475 | 6834 | 220,709 | 6157 | 745 | 4513 |
| LUSC          | 500            | 18,277       | 3,787,605        | 204,145 | 6367 | 173,856 | 11,934 | 1050 | 10,487 |
| MESO          | 87             | 17,742       | 4,904,165        | 16,527 | 475 | 16,140 | 474 | 43 | 471 |
| OV            | 301            | 18,137       | 3,018,011        | 92,743 | 7100 | 74,419 | 6196 | 2028 | 2245 |
| PAAD          | 178            | 18,021       | 5,099,858        | 113,810 | 2468 | 104,058 | 1212 | 110 | 978 |
| PRAD          | 494            | 17,646       | 4,887,130        | 691,299 | 10,152 | 514,457 | 1146 | 241 | 985 |
| READ          | 94             | 17,427       | 4,653,098        | 22,788 | 781 | 22,114 | 72 | 14 | 72 |
| SARC          | 258            | 18,183       | 4,156,361        | 70,201 | 4194 | 61,193 | 5704 | 1055 | 4115 |
| SKCM          | 103            | 17,645       | 4,908,336        | 15,046 | 720 | 14,487 | 348 | 45 | 299 |
| STAD          | 415            | 18,478       | 4,362,659        | 161,271 | 4913 | 142,709 | 2470 | 391 | 1994 |
| TGCT          | 150            | 17,790       | 4,927,197        | 71,832 | 1959 | 67,882 | 653 | 39 | 599 |
| THCA          | 503            | 17,277       | 4,936,390        | 927,678 | 10,766 | 659,323 | 13,592 | 745 | 8904 |
| THYM          | 120            | 17,785       | 5,036,992        | 85,627 | 2090 | 78,507 | 436 | 43 | 379 |
| UCCE          | 176            | 18,195       | 5,111,002        | 25,426 | 1188 | 24,721 | 251 | 35 | 248 |
| UCS           | 56             | 18,314       | 4,036,518        | 488 | 25 | 488 | 6 | 2 | 6 |
| UVM           | 80             | 16,758       | 4,812,283        | 26,233 | 890 | 25,260 | 5 | 4 | 5 |

*aThe full names of cancer types are shown in Supplementary Table S1.

Data browsing and querying of four modules

Using the homepage browser bar or clicking directly on the ‘cis/trans-eQTLs’ module, users can enter cis/trans-eQTLs page. A table with SNP ID, SNP genomic position, SNP alleles, gene symbol, gene position, beta value (effect size of SNP on gene expression) and eQTL P-value are displayed on the cis/trans-eQTLs page. When the user selects a specific cancer type or enters a gene or SNP ID, the table will be rebuilt to display the query results. For each record of SNP-gene pairs, a vector diagram of boxplot is provided to display the association between SNP genotypes and gene expression. For example, our analysis showed that ERAP2 expression in individuals carrying the homozygote rs2351010 aa is significantly higher than that in individuals carrying the homozygote rs2351010 AA and heterozygous rs2351010 Aa (P-value = 2.37 × 10^-302) (Figure 2C).

On the survival-eQTLs page, the SNP information, regulated gene information and related GWAS traits are displayed. Search boxes are designed for retrieving specific cancer types and SNPs. In addition, users can select a different LD threshold from the dropdown box to prioritize SNPs.

SUMMARY AND FUTURE DIRECTIONS

We systematically identified cis-eQTLs, trans-eQTLs, survival-associated eQTLs and GWAS-related eQTLs in 33 cancer types. We constructed a user-friendly database, PancanQTL, for users to query, browse and download eQTLs. Millions of vector diagrams of eQTL box plots and KM plots are provided. PancanQTL could serve as an important resource for human cancer genetics and provide opportunities to bridge the knowledge gap from variants in sequence to phenotypes. PancanQTL could also contribute to understanding the effects of inherited variants in tumorigenesis and development. Cancer genomics is a rapidly developing field (31), and we expect that the number of cancer samples with genotype and gene expression profiles will increase dramatically. We will update PancanQTL to include more cancer samples and will maintain it as a useful resource for the research community. Previous studies demonstrated the complicated mechanisms for regulating gene expression by eQTLs, including altering RNA
sequence, RNA structure, transcription factor binding, miRNA binding, methylation and histone modification (32,33). It will be very interesting to further investigate the regulating mechanisms of eQTLs through integrative analysis if multi-dimensional data are available.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
The authors gratefully acknowledge Xianchun Tu for helping design and debug the website, and Carol K. Kohn for proofreading of the manuscript. We thank the support from the Cancer Prevention & Research Institute of Texas (CPRIT RR150085).

FUNDING
National Natural Science Foundation of China [81402744 to J.G.]; Cancer Prevention & Research Institute of Texas [RR150085 to L.H.]; UTHealth Innovation for Cancer Prevention Research Training Program Post-doctoral Fellowship (Cancer Prevention and Research Institute of Texas) [RP160015]; China Scholarship Council [201606160058 to C.L., 201606275095 to J. F.]. Funding for open access charge: National Natural Science Foundation of China [81402744].

Conflict of interest statement. None declared.

REFERENCES
1. Wu,C., Miaox., Huang.L., Che.X., Jiang.G., Yu.D., Yang.X., Cao.G., Hu.Z., Zhou.Y. et al. (2011) Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. Nat. Genet., 44, 62–66.
2. Visscher,P.M., Wray,N.R., Zhang,Q., Sklar,P., McCarthy,M.I., Brown,M.A. and Yang.J. (2017) 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet., 101, 5–22.
3. Schork,N.J., Fallin,D. and Lanchbury,J.S. (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. Clin. Genet., 58, 250–264.
4. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicke,P., Manolio,T., Hindorff,L. et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 42, D1001–D1006.
5. MacArthur,J., Bower,E., Cezaro,M., Gil.L., Hall,P., Hastings,E., Junkins,H., McMahon,A., Milano,A., Morales,J. et al. (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 45, D896–D901.
6. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A., 106, 9362–9367.
7. Westra,H.J., Peters,M.J., Esko,T., Yaghootkar,H., Schurmann,C., Kettunen,J., Christiansen,M.W., Fairfax,B.P., Schramm,K., Powel,E. et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet., 45, 1238–1243.
8. Zhu,Z., Zhang,F., Hu,H., Bakshi,A., Robinson,M.R., Powell,J.E., Montgomery,G.W., Goddard,M.E., Wray,N.R., Visscher,P.M. et al.

Figure 2. Overview of PanCanQTL database. (A) Four modules in PanCanQTL, including cis-eQTLs, trans-eQTLs, survival-associated eQTLs and GWAS-related eQTLs. (B) Advanced search box in PanCanQTL. (C) Example of an eQTL boxplot in cis-eQTL page. (D) Example of a KM plot in survival-eQTL page.
19. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5, e1000529.

20. Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. et al. (2015) A global reference for human genetic variation. *Nature*, 526, 68–74.

21. Graffelman, J. (2015) Exploring diallelic genetic markers: the hardy weinberg package. *J. Stat. Softw.*, 64, 1–23.

22. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.

23. Kang, H.M., Ye, C. and Eskin, E. (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180, 1909–1925.

24. Leek, J.T. and Storey, J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3, 1724–1735.

25. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38, 904–909.

26. Stegle, O., Parts, L., Pietrusiak, M., Winn, J. and Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7, 500–507.

27. Shabalin, A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28, 1353–1358.

28. Gentles, A.J., Huang, S., Song, Y., Zou, Y., Westfall, P., Kim, D., Nair, V.S., Xu, Y., Hsiao, K., Ko, A., Hoang, C.D. et al. (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.*, 21, 938–945.

29. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O’Connell, R.J. and de Bakker, P.I. (2008) SNP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24, 2938–2939.

30. Gong, J., Liu, C., Liu, W., Xiang, Y., Guo, A.Y. and Han, L. (2017) LNCediting: a database for functional effects of RNA editing in lncRNAs. *Nucleic Acids Res.*, 45, D79–D84.

31. Garroway, L.A. and Lander, E.S. (2013) Lessons from the cancer genome. *Cell*, 153, 17–37.

32. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, 16, 197–212.

33. Shastry, B.S. (2009) SNPs: impact on gene function and phenotype. *Methods Mol. Biol.*, 578, 3–22.