The Author(s) *BMC Proceedings* 2016, **10**(Suppl 7):31
DOI 10.1186/s12919-016-0014-0

BMC Proceedings

PROCEEDINGS Open Access

Association of genetic variations and gene expression in a family-based study

Achilleas N. Pitsillides[1*], Seung-Hoan Choi[2], John D. Hogan[3], Jaeyoung Hong[2] and Honghuang Lin[1,4*]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Background: Expression quantitative trait locus (eQTL) maps are considered a valuable resource in studying complex diseases. The availability of gene expression data from the Genetic Analysis Workshop 19 (GAW19) provides a great opportunity to investigate the association of gene expression with genetic variants in blood.

Methods: A total of 267 samples with gene expression and whole genome sequencing data were employed in this study. We used linear mixed models with genetic random effects along with a permutation procedure to create an eQTL map. The eQTL map was further tested in terms of functional implication, including the enrichment in disease-related variants and in regulatory regions.

Results: We identified 22,869 significant eQTLs from the GAW19 data set. These eQTLs were highly enriched with genetic loci associated with blood pressure and DNase hypersensitive regions. In addition, the majority of genes associated with eQTLs showed moderate to high heritability ($h^2 > 0.4$).

Conclusions: We successfully created an eQTL map from the GAW19 data set. Our study indicated that the eQTLs were enriched within regulatory regions, and tended to have relatively high heritability.

Background

Gene expression is an essential component of the central dogma of molecular biology, and is mediated by both genetic and environmental factors [1]. Expression quantitative trait loci (eQTLs) are genomic loci that regulate gene expression. Previous studies have found that eQTLs are enriched with disease-related variants uncovered by genome-wide association studies (GWAS) [2], suggesting that eQTLs might be a good resource to better understand the functional implication of GWAS signals, and uncover potential molecular mechanisms underlying disease etiology. The availability of gene expression data in the Genetic Analysis Workshop 19 (GAW19) provided us an opportunity to investigate the association of gene expression with genetic variants identified through whole genome sequencing [3].

* Correspondence: anp4r@bu.edu; hhlin@bu.edu
[1]National Heart Lung and Blood Institute's Framingham Heart Study, 73 Mt. Wayte Avenue, Suite 2, Framingham, MA 01702, USA
Full list of author information is available at the end of the article

Methods

Data description

For each analysis we used real phenotype data from individuals in the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) whole genome sequencing (WGS) families provided by GAW19, which included systolic and diastolic blood measurements, age, sex, and smoking status. These measurements were quantified in four visits, and data for at least one visit was available for 939 participants. A full description of the data can be found in Almasy, et al. [4].

The genotype data was generated from WGS, which included 12,296,048 single-nucleotide variants (SNVs) on odd-numbered chromosomes found in at least one of 464 studied participants.

The gene expression profiled was previously described in detail by Göring, et al. [5]. Briefly, peripheral blood was collected from 1,240 individuals. However, only 267 samples also had their genotype profiled by WGS and thus were chosen for our current study. The RNA was

extracted from mononucleated cells, amplified, and then hybridized to an Illumina Sentrix Human Whole Genome (WG-6) microarray, which measured the expression profiles of more than 20,000 transcripts. The expression data were collected at the first visit, so we only use trait data and covariates from that visit. Because only 34 individuals used blood pressure medication and we did not expect different types of medication to have the same effect on gene expression, we did not adjust for blood pressure medication use.

Annotation

The functional annotation of genetic variants was performed using the ANNOVAR (Annotate Variation) package [6]. We defined putative *cis*-eQTLs as SNPs located within 1 Mb of the transcript starting sites, and *trans*-eQTLs as SNPs located more than 1 Mb of the transcript starting sites, or in different chromosomes.

Expression quantitative trait locus analysis

The association of genetic variants with gene expression (eQTL) was tested by linear mixed effects models, adjusting for random genetic effects and age, sex, smoking status and the first two components from the expression principal components analysis ($Pc1$, $Pc2$) as covariates. Using the subscript ij to denote the j^{th} individual in the i^{th} family, and defining Y_{ij} and SNP_{ij} as the gene expression and genotype dosage, respectively, we write the model as:

$$Y_{ij} = \beta_o + \beta_1 age_{ij} + \beta_2 sex_{ij} + \beta_3 smoking_{ij} + \beta_4 Pc1_{ij} + \beta_5 Pc2_{ij} + \beta_s SNP_{ij} + \alpha_{ij} + \varepsilon_{ij}$$

where the betas denote the regression coefficients for the fixed effects, α_{ij} is the random intercept, and ε_{ij} the normally distributed error term with mean zero and variance σ^2_ε. The α_{ij} within the i^{th} family are normally distributed with mean zero and covariance matrix $\sigma^2\Phi$, where Φ is the kinship matrix. The overall covariance matrix is block diagonal with one block per family. The model was fitted in R using the *lmekin* function from the *coxme* package [7]. The inclusion of the principal components in the model is to account for batch effects.

Because of small sample size (N = 267), we had limited power to uncover *trans*-eQTLs. In particular, with our pedigree structure, assuming gene expression is fully heritable, we have 30 % power to detect a common SNP that would explain 0.2 of the variance at a significance level of $\alpha = 1E-8$. Consequently, this study concentrated on *cis*-eQTLs, which were defined as SNPs within 1 Mb from the start or end of the transcript [8].

We ran a Shapiro-Wilk normality test for all the expression probes and we found that the test was significant after a Bonferroni correction for more than half of the probes. To reduce type I error that may arise from violating the normality assumption, we estimated significance using a modified version of the permutation procedure proposed by Iturria, et al. [9] that preserves familial correlation. First, we estimated the heritability for each transcript and binned the transcripts according to the estimated heritability. For each transcript, we picked a surrogate from the same heritability bin to represent the expected association that we would have observed with a random transcript unassociated with the SNP of interest. We then reordered the original transcript values according to the ranking of the surrogate gene values. We reran the analysis with the permuted phenotype values and the putative *cis*-eQTL for the original transcript. We repeated the above procedure for each transcript three times to obtain a p value between *cis*-eQTL and transcript.

To obtain the significant results we split the genes into seven heritability bins. This splitting choice gave an approximately equal number of permutations per bin. To adjust for the different number of SNPs tested (N_g) for each gene and account for linkage disequilibrium, we multiplied the p values by the factor $\frac{N_g+1}{2}$, which has been shown to be a good estimate of the number of effective tests [10]. We used the minimum adjusted p value from each permutation to form a null distribution of the adjusted p value statistic for each heritability bin.

Enrichment of eQTLs in disease-related variants and regulatory regions

We took all the significant eQTLs and performed a Fisher's exact test to determine if the eQTLs were significantly enriched with known genetic loci associated with blood pressure. A total of 26 top SNPs were collected from the literature. However, none of these SNPs were eQTLs found in this study, either because they were not directly sequenced, or because they were rare variants (minor allele frequency (MAF) <5 %). Therefore, we extended this list to nearby SNPs with distance less than 1 kb and r^2greater than 0.9. This extended list contained 31 SNPs that were also eQTLs in our study.

We also examined if eQTLs were significantly enriched within DNase hypersensitive regions based on ENCODE data [11].

Results

Expression quantitative trait locus map

Here, we report significant expression SNPs (eSNPs) controlling for a false discovery rate of 0.05 using the Benjamini-Hochberg procedure [12]. We found 21,753 significant eSNPs that controlled the expression of 332 distinct Genes (eGenes); 985 eSNPs targeted more than one eGene. Only one eGene has low

heritability ($h^2 < 0.1$), and all the other eGenes have medium to high heritability ($h^2 > 0.4$). The ten most significant eQTLs are shown in Table 1 and the breakdown of the results by heritability bin is shown in Table 2.

Enrichment of expression quantitative trait loci in blood pressure–related variants

To demonstrate the usefulness of our eQTL map, we examined the distribution of blood pressure–related genetic variants from GWAS found in the literature. Our variant list contained 31 SNPs that were a part of the eQTL study, two of which were eSNPs. Even with the small numbers, a Fisher's exact test (p <6.1E-4) showed that the set of GWAS results contained more eSNPs than we would expect by chance.

Enrichment of expression quantitative trait loci in regulatory regions

We found 386,135 variants that were part of the eQTL analysis that lie within DNase hypersensitivity regions, of which 5,679 were found to be eSNPs. Our results confirmed the enrichment of eQTLs in DNase cluster regions (Fisher's exact test p value <2.2e-16), which contained 17 times more eQTLs than random variants.

Association of gene expression with blood pressure

We then tested the association of gene expression with blood pressure, and found that two probes were significantly associated with blood pressure. One probe (GI_7706275.A with p <5.5E-9) maps to the gene *TPP3* on chromosome 16, while the other (GI_42661149 with p <1.8E-6) maps to the predicted gene *LOC400604*.

Discussion

The identification of genetic mechanisms underlying gene expression would enable a better understanding of the biology of complex diseases. We performed a comprehensive analysis of the association between gene expression and genetic variants using GAW19 data, and successfully identified 22,869 eQTLs despite limited sample size. These eQTLs were highly enriched in genetic loci associated with blood pressure and DNase hypersensitive regions, suggesting that eQTLs might play an important role in regulatory functions. Our breakdown of eSNP results by heritability indicates that most of these eSNPs have high heritability. Future investigation with a larger sample size would further validate the association of heritability with eQTLs.

Table 1 Top 10 eQTLs

Gene symbol	Chr	Pos	Ref	Alt	Raw *p* value	Adjusted *p* value
UBA52	19	18668135	G	C	<1e-324	<1e-324
UTS2	1	7970383	T	C	<1e-324	<1e-324
CAPZA1	1	113138318	G	A	1.9e-150	3.0e-147
TMEM176B	7	150478052	G	T	1.9e-139	3.2e-136
TMEM176A	7	150460891	T	G	1.7e-134	3.0e-131
RPL14	3	40498845	T	C	3.3e-123	4.4e-120
TIMM10	11	57324428	G	A	1.2e-120	1.9e-117
SPATA20	17	48624523	A	C	1.8e-112	2.7e-109
SLC12A7	5	1104938	C	T	4.0e-97	7.7e-94
RNF167	17	4848450	G	A	1.1e-92	2.2e-89
RPS6KB2	11	67197757	G	A	3.9e-90	2.9e-87

Table 2 Summary results by h^2 bin

h^2	eSNPs	Genes tested	SNPs per genes (mean)
0–0.1	4	451	3253
0.1–0.2	0	580	3111
0.2–0.3	0	800	3198
0.3–0.4	0	1145	3141
0.4–0.5	2232	1272	3210
0.5–0.6	350	1079	3123
0.6–1	20283	904	3177

Conclusions

We created an eQTL map that included 22,869 eQTLs from GAW19 data. We found that eQTLs were enriched within regulatory regions, and tended to have relatively high heritability.

Acknowledgements

All the analyses were performed on the Boston University Shared Computing Cluster.

Declarations

This article has been published as part of *BMC Proceedings* Volume 10 Supplement 7, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at http://bmcproc.biomedcentral.com/articles/supplements/volume-10-supplement-7. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Authors' contributions

ANP drafted the manuscript, ran enrichment analysis, and obtained the eQTL analysis significant results. SHC performed the permutation based eQTL analysis and edited the manuscript. JDH collected SNPs from the literature and edited the manuscript. JH obtained the annotation information of genes and edited the manuscript. HL conceived the study, oversaw the analyses, and edited the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare they have no competing interests.

Author details

[1]National Heart Lung and Blood Institute's Framingham Heart Study, 73 Mt. Wayte Avenue, Suite 2, Framingham, MA 01702, USA. [2]Department of Biostatistics, Boston University School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. [3]Program in Bioinformatics, Boston

University, Boston, MA 02215, USA. [4]Department of Medicine, Boston University School of Medicine, 72 East Concord St, Boston, MA 02118, USA.

Published: 18 October 2016

References

1. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, et al. Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet. 2006;2(10), e172.
2. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6(4), e1000888.
3. Li L, Kabesch M, Bouzigon E, Demenais F, Farrall M, Moffatt MF, Lin X, Liang L. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. Front Genet. 2013;4:103.
4. Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, Johnson M, Peralta JM, Manning AK, Wood AR, et al. Omics squared: Human genomic, transcriptomic, and phenotypic data for Genetic Analysis Workshop 19. BMC Proc. 2015;9 Suppl 8:S2.
5. Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet. 2007;39(10):1208–16.
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16), e164.
7. Therneau T. coxme: Mixed effects Cox models. *R package version* 2.3. 2012, https://cran.r-project.org/web/packages/coxme/index.html.
8. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6(5), e107.
9. Iturria SJ, Williams JT, Almasy L, Dyer TD, Blangero J. An empirical test of the significance of an observed quantitative trait locus effect that preserves additive genetic variation. Genet Epidemiol. 1999;17 Suppl 1:S169–73.
10. Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau O, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet. 2007;16(1):36–49.
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300.