Mapping Novel Pathways in Cardiovascular Disease Using eQTL Data: The Past, Present, and Future of Gene Expression Analysis

Citation

Published Version
doi:10.3389/fgene.2012.00232

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11708606

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Mapping novel pathways in cardiovascular disease using eQTL data: the past, present, and future of gene expression analysis

Rajat M. Gupta1,2* and Kiran Musunuru1,2

1 Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA
2 Division of Cardiovascular Medicine, Brigham and Women’s Hospital, Boston, MA, USA

Edited by:
Liming Liang, Harvard School of Public Health, USA

Reviewed by:
Wei-Min Chen, University of Virginia, USA
Dajiang Liu, University of Michigan, USA

*Correspondence:
Rajat M. Gupta, Division of Cardiovascular Medicine, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02115, USA.
e-mail: rajatmgupta@gmail.com

Genome-wide association studies (GWAS) have identified genetic variants associated with numerous cardiovascular and metabolic diseases. Newly identified polymorphisms associated with myocardial infarction, dyslipidemia, hypertension, diabetes, and insulin resistance suggest novel mechanistic pathways that underlie these and other complex diseases. Working out the connections between the polymorphisms identified in GWAS and their biological mechanisms has been especially challenging given the number of non-coding variants identified thus far. In this review, we discuss the utility of expression quantitative trait locus (eQTL) databases in the study of non-coding variants with respect to cardiovascular and metabolic phenotypes. Recent successes in using eQTL data to link variants with functional candidate genes will be reviewed, and the shortcomings of this approach will be outlined. Finally, we discuss the emerging next generation of eQTL studies that take advantage of the ability to generate induced pluripotent stem cell lines from population cohorts.

Keywords: cardiovascular disease, eQTL, GWAS, induced pluripotent stem cells, RNA-seq

eQTL ANALYSES AND BLOOD CHOLESTEROL LEVELS

A recent study used eQTL data to elucidate the relationship between a non-coding GWAS variant in a chromosome 1p13.3 locus associated with low-density lipoprotein (LDL) cholesterol and the hepatic expression of the SORT1 gene (Musunuru et al., 2010). To determine if SNPs in the 1p13.3 locus were cis-acting regulators of nearby genes, expression levels were measured in liver, subcutaneous fat, and omental fat human tissue samples. After a large-scale analysis showed association between a representative SNP (rs646776) and transcript levels of three genes (CELSR2, PSRC1, and SORT1) in liver, a targeted eQTL analysis was performed in a replication cohort of 62 human liver samples. Minor allele homozygotes displayed more than 12-fold higher SORT1 expression than major allele homozygotes. This differential gene expression by genotype was not apparent in the analysis of the adipose samples or in previously reported data from lymphocytes (Linsel-Nitschke et al., 2010), suggesting that the mechanism underlying the allele-specific gene expression is liver-specific. This finding paved the way for in vivo experiments modulating liver gene expression with siRNA or viral overexpression vectors, which established that Sort1 alters blood LDL cholesterol levels by modulating hepatic lipoprotein particle secretion (Musunuru et al., 2011; Min et al., 2011). Furthermore, it has been shown that many common trait-associated variants act by altering transcript levels (Nicolae et al., 2010). Thus the use of eQTL data can provide the link between a SNP of interest and its responsible coding gene, and it is a reasonable initial line of inquiry in the search for functionally significant SNPs.
Although there are other genes at the loci in question, strong eQTL evidence that reached genome-wide significance with monocyte transcripts (RNA-seq) were found to be due to the separate processing of lymphoblastoid cell lines (LCLs) from subjects of European vs. Asian ancestry. From monocyte-derived eQTL data, there appear to be cis relationships with nearby genes CDKN2A, CDKN2B, and ANRIL (Cunnington et al., 2010). But these data do little to elucidate the mechanism by which the 9p21 locus contributes to disease. For a biologic connection between a SNP and a mechanism of disease much more experimentation is required; though eQTL analysis may provide the initial insight.

IDENTIFIED eQTLs DO NOT ALWAYS LEAD TO DIRECT BIOLOGICAL INTERPRETATION

Single nucleotide polymorphisms in the CAD-associated chromosome 9p21.3 locus – by far the strongest population-wide genetic contributor to CAD – have evaded mechanistic understanding though they are associated with not just CAD but also aneurysms, vascular disease, and multiple cancers (Helgadottir et al., 2008). From monocyte-derived eQTL data, there appear to be cis relationships with nearby genes CDKN2A, CDKN2B, and ANRIL (Cunnington et al., 2010). But these data do little to elucidate the mechanism by which the 9p21 locus contributes to disease. For a biologic connection between a SNP and a mechanism of disease much more experimentation is required; though eQTL analysis may provide the initial insight.

TECHNICAL LIMITATIONS: PLATFORM COVERAGE, BATCH VARIATION, SAMPLE SIZE, AND TISSUE AVAILABILITY

There currently exist several commercial platforms by which expression data is obtained. The coverage of each platform is variable, and each is subject to its individual errors. This variability results in batch-to-batch variation between analyses that should otherwise be consistent. This was observed in a report of large-scale differences in gene expression between ethnic groups (Akey et al., 2007). In this case, the highly significant differences were found to be due to the separate processing of lymphoblastoid cell lines (LCLs) from subjects of European vs. Asian ancestry. Different microarray platforms only have 30–40% overlap in transcript detection (Barnes et al., 2005; Pedotti et al., 2008). With the advent of direct ultra-high-throughput sequencing of RNA transcripts (RNA-seq), analyses of gene expression are expected to improve; however, handling the terabytes of data generated by such approaches will be a challenge in its own right.

Suggested solutions for batch-to-batch variation include experimental and statistical approaches. Efforts should be made to spread sample groups across different processing times to limit this as a confounding variable. To correct for batch effects with statistical analyses, the possible variable that could contribute should be identified and reported. Published results, for example, should report the processing group and time of samples in a study (Leck et al., 2010).
Many eQTL databases are limited to a few 100 human samples because of cost and difficulty in obtaining tissue. This results in nominal P values in most eQTL analyses that are difficult to interpret (Schadt et al., 2003; Morley et al., 2004). To rectify this limitation, the National Institutes of Health have launched a 2-year pilot project to build a more comprehensive bank of samples termed the Genotype-Tissue Expression project (GTEx). To date, no project has analyzed genetic variation and expression in as many tissues in such a large cohort as planned for GTEx. If completed, GTEx will house samples of 30 different tissues from 1,000 donors. Even with this large database, however, questions remain such as whether healthy or diseased samples should be collected together (Cookson et al., 2009).

The fact that many eQTL studies are conducted in LCLs is both a technical and conceptual shortcoming. LCLs form the basis for numerous gene expression studies because they represent an easily banked source of nucleic acids for genetic studies (Cookson et al., 2009). LCLs, however, exhibit genomic instability with multiple passages of storage and regrowth. Furthermore, LCLs likely have unique, tissue-specific expression profiles that may not reflect normal human biology. The predominance of clonal cell populations, for example, can result in random patterns of monoallelic expression (Plagnol et al., 2008). There is recent evidence, however, that other cell types have expression overlap with LCLs. In fact, 70% of cis-eQTLs in LCLs are shared with skin cells (Ding et al., 2010). This finding can allow for comparisons between tissue types, and conclusions about the functional importance of variants can potentially be made on this basis.

Even when eQTL studies use cells other than LCLs, there are inherent biases as a result of the limitations of tissue ascertainment. Many of the studies that profile liver and adipose expression use human samples obtained during surgeries (Schadt et al., 2008); it stands to reason that these were from non-healthy patients undergoing surgery for non-research-related reasons, since it would be unethical to perform such biopsies in healthy individuals. Thus the samples do not represent the general population, nor could they be ascertained a priori on the basis of particular genotypes or phenotypes of interest. Furthermore, the samples were limited in size, non-renewable, and could not be used to generate large numbers of cultured hepatocytes and adipocytes, since unlike LCLs primary hepatocytes and adipocytes cannot be sustained for long in tissue culture conditions.

**FUTURE DIRECTIONS**

The full spectrum of gene expression that relates to disease involves not only many cell types but different conditions for these cells to “exercise the genome.” There is evidence that environmental actions on gene expression are profound in humans, and possible, future eQTL studies should incorporate environmental stimuli. Model stimuli that could be tested in an in vitro system include pro-inflammatory stresses, metabolic stresses (such as hypoglycemia or hypoxia), the response to radiation, and even response to drugs, hormones, and peptides (Cookson et al., 2009).

The next generation of eQTL studies will have the ability to utilize induced pluripotent stem cells (iPSCs) as a renewable source of patient-specific cell lines. Recent advances in nuclear reprogramming technology allow for the transformation of terminally differentiated adult cells into induced iPSs that are phenotypically indistinguishable from embryonic stem cells (Takahashi et al., 2007). This leap forward makes possible the creation of patient-specific iPSC lines that, by definition, can be maintained in culture indefinitely. Gene expression profiling from these patient-specific cell lines will allow larger-scale gene analyses from patients in whom complete phenotype data is available.

As one example, an effort is now underway to create iPSCs from blood samples from up to 3,000 participants in the Framingham Heart Study (FHS), which is a unique, community-based cohort in which three generations of individuals have been extensively phenotyped and genotyped. Although limited eQTL analyses have been performed within the FHS (Levy et al., 2009; Fox et al., 2011), these analyses suffer from the standard pitfalls of underpowered results and suboptimal tissue types. With iPSCs, the full spectrum of gene expression can be extensively profiled in any tissue that can be differentiated from the cell lines, allowing investigators to study tissue types that in ordinary circumstances would be prohibitive to obtain from living human beings. In the planned pilot studies, a subset of the FHS iPSC lines will be differentiated into hepatocytes and adipocytes — two tissues of relevance to cardiovascular and metabolic diseases — followed by whole-genome gene expression profiling, which will pave the way for eQTL analyses of unprecedented size and rigor.

**CONCLUSION**

Gene expression analyses can yield important information about genetic architecture and can point to mechanisms that link genetics and cardiovascular disease. The application of eQTL analyses has already forged connections between GWAS data and mechanistic pathways. Technical and conceptual limitations, however, limit the ability to systemize this approach to more polymorphisms for more diseases. The advent of more comprehensive biobanks, high-throughput RNA sequencing, and collections of iPSC lines holds much promise for the field. Ultimately, the ability to explain the heritability of common diseases such as cardiovascular and metabolic disorders should be greatly facilitated by future eQTL efforts.

**REFERENCES**


Nicolaou, D. L., Gamañon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs annotation to enhance discovery from GWAS. PLoS Genet. 6:e1000888. doi:10.1371/journal.pgen.1000888