Published in final edited form as:

Curr Opin Immunol. 2012 October ; 24(5): 544–551. doi:10.1016/j.coi.2012.09.002.

Coordinating GWAS results with gene expression in a systems immunologic paradigm in autoimmunity

Barbara E. Stranger^{a,b,c} and Phillip L. De Jager^{b,c,d}

^aDivision of Genetics, Brigham and Women's Hospital, Boston, MA, 02115, USA

^bHarvard Medical School, Boston, MA 02115, USA

^cProgram in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, 02139, USA

^dProgram in Translational NeuroPsychiatric Genomics, Department of Neurology, Brigham & Women's Hospital, Boston, MA 02115, USA

Abstract

There has been considerable progress in our understanding of the genetic architecture of susceptibility to inflammatory diseases in recent years: several hundred susceptibility loci have been discovered in genome-wide association studies (GWAS) of human populations. This success has created an important challenge in identifying the functional consequences of these risk-associated variants and in elucidating how the repercussions of individual susceptibility loci integrate to yield dysregulation of immune pathways and, ultimately, syndromic clinical phenotypes. The integration of GWAS association signals with high-resolution transcriptome and other genomic data that capture the dynamics of cellular state and function in the context of individual's collection of susceptibility alleles has proven to be a successful avenue of investigation. The rapid pace of methodological development in this area has been coupled with an accumulation of experimental data that makes the elucidation of complex biological networks underlying susceptibility to these common inflammatory diseases a reasonable goal in the near future.

Introduction

Advances in genotyping technology, together with the discovery that genetic variation in the human genome is structured in such a way that common nucleotide variants do not segregate entirely independently of each other, ushered in an era of genome-wide association studies (GWAS). A GWAS is now a common study design for discovering genetic variation that contributes to complex traits and, to be successful, typically requires the evaluation of hundreds of thousands to millions of genetic variants for correlation to a given phenotype in several thousand individuals. To date, approximately 200 loci harboring commonly-occurring genetic variants (minor allele frequency > 0.05) associated with disease risk have been convincingly associated with inflammatory disease in humans, and the National

© 2012 Elsevier Ltd. All rights reserved

Corresponding author: Barbara E. Stranger Brigham and Women's Hospital Department of Medicine, Division of Genetics 77 Avenue Louis Pasteur New Research Building 168B Boston, MA 02115, USA Phone: +1 617-525-4767 Fax: +1 617-525-5333
bstranger@rics.bwh.harvard.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Human Genome Research Institute at the U.S. National Institutes of Health updates a catalog of published GWAS results (<http://www.genome.gov/gwastudies/>) on a weekly basis. Yet, despite our success in discovering susceptibility loci, the identification of the precise variant(s) contributing to a given trait and the mechanism by which such variants exert their effect on disease remain elusive. One challenge is the sheer number of susceptibility loci, as the detailed dissection of a single locus represents a substantial investment in effort and resources. Typically, many loci are associated with trait variance, and each locus contributes only a small effect to syndromic traits such as susceptibility to an inflammatory disease. Thus, *a priori*, there is often not a clear order with which to proceed.

Another challenge is that a locus, the segment of chromosomal DNA containing the trait-associated variant(s), may contain multiple genes: mapped trait-associated variants localize to gene-rich as well as gene-poor regions. Importantly, most of the associated variants within a given locus are surrogate markers that are in linkage disequilibrium (LD) not only with the causal variant(s) but also many other variants. Thus, the causal variants are not readily identified at the end of a genome scan. Nonetheless, this set of correlated variants is useful in that it defines the boundaries of the locus that contains the causal variant(s). The gene(s) present in this chromosomal segment are the ones that are most likely to be affected by the disease-associated variant (although long distance regulatory effects are also possible), but, if there is more than one gene in the region, often one cannot statistically differentiate which one may be more likely to be affected. Since only a small number of trait-associated variants are coding variants that affect protein sequence non-synonymously and have been proven to have an effect on gene function, most loci require fine mapping of an association and further characterization of a locus to understand their role in the trait of interest.

Since a causal chain links a risk factor such as a genetic variant to immune dysfunction and eventually a syndromic phenotype such as susceptibility to an inflammatory disease, identifying the effect of variants in a susceptibility locus on pertinent intermediate phenotypes has proven to be a fruitful strategy with which to explore the functional consequences of a susceptibility locus and to refine the identity of the causal allele. Gene expression is one such intermediate trait that has been successfully leveraged in a number of disease studies [e.g., 1–7]. In this article, we review the current state-of-the art for integrating gene regulatory genomics with GWAS results in a systematic manner, using studies of inflammatory disease variants to illustrate the different strategies that have been successfully deployed.

The genetic basis of gene regulatory variation

While DNA sequence variants, such as null alleles, can result in extremes of gene expression that may be deleterious, population-based studies of healthy individuals have reported high levels of heritable inter-individual variation in gene expression levels, and studies have mapped genetic variation contributing to gene expression levels in a number of different cell types [8–14]. Collectively, studies that map genetic variation contributing to transcriptional variation are referred to as expression quantitative trait locus (eQTL) mapping studies. Their general design consists of genotyping subjects genome-wide and capturing a transcriptome-wide mRNA profile using microarrays, or more recently high-throughput RNA-sequencing. As in gene discovery studies related to a given disease, imputation using a reference map of human genome variation is used to interrogate the role of variants that are not directly genotyped but are in LD with a genotyped marker. Thus, almost every common marker (e.g., those alleles with a minor allele frequency > 5%) evaluated in GWAS studies will have been evaluated, directly or indirectly, in an eQTL study. An eQTL analysis itself consists of applying regression-based or non-parametric

models to test millions of genetic variants for regulatory effects on the expression of nearby and distant genes. In a whole-genome eQTL analysis, many millions of tests are performed ($N = \text{number of genetic variants} \times \text{number of genes or transcripts}$), requiring strict statistical thresholds for significance. While necessary in a genome-wide analysis, such strict thresholds can obscure many biologically meaningful effects of genetic variation on mRNA expression. Thus, approaches to appropriately constrain such analyses have evolved: based on our understanding of the architecture of mammalian genes, sequences involved in the regulation of gene expression of a given gene are most likely to be found near that gene and harbor genetic variation influencing gene expression. Such '*cis*-eQTL' analyses are focused on assessing the role of genetic variants on the expression of genes in their vicinity and, empirically, have been demonstrated to be well-powered to detect regulatory effects that are replicated [15]. For a given tissue and cohort, these analyses provide a list of genetic variants associated with a given gene or transcript's expression levels, the allelic direction of the association, and the magnitude of the effect, often quantified as the mean change in expression between individuals homozygous for either of the two alternate alleles. As such, they provide annotation of functional regulatory variation in the human genome.

When GWAS variants and eQTL variants co-localize to the same genomic region, this generates a testable hypothesis that a given genetic variant influences trait variance through effects on expression of a given gene. An example of identification of an eQTL that co-localizes with Multiple Sclerosis (MS) associated variants is shown in Figure 1. The Figure shows a region of chromosome 20 where a genetic variant contributing to MS susceptibility has been localized (lower panel). This susceptibility locus encompasses multiple genes; however, the MS-associated genetic variant and surrogate markers that are in LD with it are also associated with altered gene expression of one of the genes in the region, specifically the CD40 gene, in peripheral blood mononuclear cells of multiple sclerosis patients. This result suggests that MS susceptibility may be, in part, due to altered CD40 gene expression levels that are influenced by genetic variants near to the gene. The result is by no means conclusive, but, it does suggest testable hypotheses to pursue in future studies.

To date, the majority of eQTL studies have been performed in healthy subjects and reveal a substantial amount of functional regulatory variation in the human genome. The degree to which the detected associations are population- and cell-type specific, or observable only under certain conditions, is an important consideration for integrating with GWAS associations. Clearly, some genes exhibit highly cell type-, tissue-, and context-specific expression patterns [16–17], and the extent to which the eQTL patterns are shared across cell-types or tissues is still being quantified. Most eQTL studies to date have been of modest size, limiting the assessment of tissue overlap because of poor statistical power to support a negative result. Published studies have estimated that only 0.4%–0.5% of genes have a significant *cis*-eQTL in at least two or more tissues, and approximately 70–80% of eQTLs may be cell-type specific [8–10,18–21]. While these modestly sized studies may overestimate the true proportion of cell-type specific eQTLs, it is clear that, in many situations, the cell- or tissue-type in which mRNA is profiled will have an important effect on the presence and magnitude of a variant's effect on gene expression. This is important as it determines the relevance of existing eQTL datasets to disease studies and informs the design of new studies that have to balance the competing needs of accessing a specific cell type of interest (which may be technically and practically challenging) or using a suitable surrogate cell population or cell mixture. To date, the majority of human eQTL studies have been conducted using cell lines, e.g., lymphoblastic cell lines (LCLs) that are derived from B lymphocytes transformed using the Epstein-Barr virus [13–14,22–24], but increasingly studies are being conducted in primary cells or tissues, including blood [3,6,19,25–26], monocytes [9,12], primary B-cells[9], liver [11,27–29], skin[10,18], adipose tissues[3,30], skeletal muscle [31], brain [32–34], T-cells [8,35], and fibroblasts [8]. Recognizing the

utility of eQTL analysis and the tissue specificity of eQTL associations, The U.S. National Institutes of Health has funded an unprecedented multi-tissue eQTL study of human gene expression (<http://commonfund.nih.gov/GTEX/>). The Genotype-Tissue Expression (GTEx) program will publish summarized eQTL analysis results from many human tissues and will catalyze methods development to best utilize this wealth of data.

Population ancestry also contributes to context specificity of eQTLs. A *cis*-eQTL analysis in LCLs of eight cohorts representing different human ancestries estimates that up to 31% of well-annotated genes have a significant *cis*-eQTL relationship in at least one population. Of these genes with *cis*-eQTLs, more than 50% exhibit that *cis*-eQTL in at least two independent populations, and 6% of genes contain an eQTL effect in all eight populations [13]. This pattern likely reflects the modest sample size of each population, effects of the transformation process, possible effects of environmental variables on the source tissue, and some clear examples of population-specific associations. This suggests that in some cases, population ancestry may be an important consideration for integration of eQTL observations with GWAS results, particularly as GWAS is applied to populations that are not of European ancestry. Interestingly, while variation in eQTL relationships among populations is expected because of differences in LD structure among human populations, current studies suggest that allele direction and the magnitude of the effect on expression levels vary little for those associations that are shared across human populations [13]. Thus, in a minority of loci, there may be little in the way of population-specific modifying effects, and further work is needed to understand the role of these eQTLs that may relate to fundamental aspects of human biology. It is important to note that nearly all eQTL investigations conducted to date have focused on common genetic variants (minor allele frequency > 5%), thus our understanding of the levels and patterns of regulatory variation and effect sizes are limited to this set. However, with recent developments in genome sequencing and rare variant genotyping and imputation, investigators are now beginning to identify and characterize regulatory effects of low-frequency variants [36], which may help in interpreting rare variant trait associations that are identified by GWAS or whole-genome sequencing projects.

Studies have demonstrated that, as a group, trait-associated variants identified through GWAS are enriched for eQTLs [37–38], with inflammatory disease risk variants showing a similar degree of enrichment for eQTLs in relevant cell types [6,25,39–41]. Publicly-available eQTL genome browsers and databases (Table 1) provide useful resources for investigating potential eQTL associations for SNPs of interest. Though, because most report only significant findings, and analyses use different statistical models and criteria for data inclusion, false negatives are difficult to interpret. Even in the case where a trait-associated variant is also determined to be an eQTL, one has to exercise caution in interpretation. Due to patterns of LD in the human genome and the large number of disease-associated and eQTL-annotated genetic variants, a simple overlap of significant eQTL variants and GWAS variants can occur even when causal variants underlying both signals are different. Statistical frameworks have now been developed to provide an estimate of the degree to which these signals of association overlap [41–42], and further methods development continues in this area.

Functional experiments to elucidate mechanisms underlying disease associations need to be performed in relevant cell types due to the context specificity of many cellular processes. Until recently, identifying the most relevant cell type for a trait of interest has relied on educated guesses based on interpretation of current literature. For inflammatory diseases, this poses a particular challenge: while a specific cell type may have a prominent role based on our current understanding of disease pathophysiology and treatment mechanism, inflammatory diseases are systemic diseases and multiple different immune cells are likely to contribute to disease susceptibility and to be the target of susceptibility variants. Studies

quantifying enrichment of eQTLs among inflammatory disease-associated variants have identified specific cell types where the enrichment is greatest and have suggested that these cell types may be the ones that are more relevant in terms of genetic susceptibility to an inflammatory disease [9,38]. A different approach has been proposed by Raychaudhuri and colleagues [43] who devised a systematic assessment of cell types that starts with a set of disease-associated variants, identifies genes within each locus, and calculates a probabilistic model that designates the tissue or cell type(s) in which the expression of these potential susceptibility genes is enriched. Applying this method to three inflammatory diseases and using RNA expression profiles from a large number of purified murine immune cells from the Immunological Genome Consortium [44], they designate cell types that are most relevant to systemic lupus erythematosus, Crohn's disease, and rheumatoid arthritis. While limited by the completeness of the list of susceptibility alleles and the use of gene sets defined in murine cells, these results implicate different combinations of cell types for each disease that are consistent with our current understanding of pathophysiology. Similar approaches using genome-wide maps outlining the state of chromatin in different human cell types have also successfully identified cell types relevant to specific diseases and will help to design studies that seek to explore disease-related eQTLs (Raychaudhuri, *personal communication*).

Systems biology to elucidate causal networks

While eQTL studies provide excellent annotation for the detailed characterization of a given locus, this becomes cumbersome when one explores the coordinated functional consequences of multiple different susceptibility loci. With each inflammatory disease having many dozens of susceptibility loci, the need for systematic and semi-automated evaluations of variant function has become acute and has led to the development of several different approaches. These approaches rely on the assumption that a set of disease-associated loci is likely to reflect a limited number of underlying mechanisms that may be detectable through enrichment analysis within biologically relevant gene sets defined using the correlation structure observed in RNA data (Kyoto Encyclopedia of Genes and Genomes; KEGG [45], Ingenuity Pathway Analysis (Ingenuity Systems), Gene Ontology; GO [46], Protein ANalysis THrough Evolutionary Relationships; PANTHER [47], Biocarta (<http://www.biocarta.com>), Reactome [48]. Enrichment analyses need to be performed carefully, taking into account parameters such as LD, gene size, pathway size, pathway complexity, etc that can easily skew results and return spurious associations that are driven by the properties of a gene set and not a true connection with disease susceptibility. Currently, INRICH [Interval-based Enrichment Analysis Tool for Genome Wide Association Studies 49]), MAGENTA [Meta-Analysis Gene-set Enrichment of variaNT Associations 50]), and DAVID [Database for Annotation, Visualization and Integrated Discovery 51,52]) are user-friendly analysis tools that conduct pathway analysis on several types of genomic variation, including but not limited to SNPs, copy number variants (CNVs), genes, as well as their combinations, taking into account several potential confounders. A different approach that uses annotations of protein-protein interactions as well as gene expression, DAPPLE [53], begins to integrate additional levels of information with RNA data and, using known susceptibility genes as a seed, has demonstrated an ability to enrich its results for genes that are later validated to have disease-associated variants [53].

Clearly, the path forward is to increase, in a meaningful way, the number and types of data considered in such analyses: the convergence of data from several different cell types identified as pertinent for a given disease may, for example, offer a better perspective on the functional repercussions of groups of variants on the interactions of different immune cells that ultimately lead to immune dysregulation and disease: certain groups of susceptibility variants may function in different cell types and have functional consequences at the cellular

level that interact with those of another group of variants in a different cell type. This type of approach gradually constrains the complexity of the analysis to build a hierarchical model outlining a network of cell-specific networks, each of which may be driven by a different subset of susceptibility variants. Such models provide an excellent substrate for the design of human immunologic studies that can validate the model and use it to investigate novel questions and perhaps develop algorithms that are useful in a clinical setting to support a diagnostic work-up.

Challenges and future directions

Integration of eQTLs with disease-associated variants is merely one tool available to try to elucidate perturbed biology and mechanisms. It is the first step in testing a very specific hypothesis, i.e., that associated variants exert their effects through gene regulatory mechanisms. If a GWAS variant is an eQTL variant in a relevant tissue or cell-type and the association signals from both data types are highly correlated, additional experiments are warranted to prove the links between genetic variation, mechanism, gene, and trait. This approach is particularly well suited to inflammatory disease where peripheral blood cell populations are relevant to the disease pathophysiology and can be sampled to rapidly confirm an eQTL observation and to develop new experiments that are based on the observation of differential gene expression relative to the susceptibility variant: the eQTL therefore provides a critical first observation that enables the rapid development of testable hypotheses. An important challenge of such functional characterization is access to healthy subjects with the common susceptibility variants of interest that can be recalled based on genotype to interrogate the role of given variants without the confounding effects of treatment and fluctuations in disease activity that are seen in subjects with an inflammatory disease. However, several resources – such as the PhenoGenetic Project at Brigham & Women's Hospital and the Cambridge BioResource – have emerged to successfully fulfill this need [54–55].

While eQTL analysis is a powerful tool, it can be enriched by adding other forms of information that also capture genetic variation that influence cellular function, such as signatures of natural selection: we and others have leveraged both types of data to gain insights into the consequences of disease-associated variants [56]. This strategy may be particularly well-suited to inflammatory disease variants since variation in immune response has clearly been under natural selection by pathogens at various times over the course of human history. Ultimately, multi-dimensional “omics” data (transcriptomics, epigenomics, lipidomics, proteomics, glycomics, etc.) from different cell-types of well-characterized cohorts, together with analytic tools to integrate varied types of data, will help to elucidate disease pathways and the manner in which genetic variants contribute to trait such as susceptibility to an inflammatory disease.

Can we already translate some of these findings in the clinical sphere? The past five years have seen a tremendous growth in our understanding of the genetic architecture of inflammatory diseases and, particularly, in the extent of their shared architecture [57]. Further, we see that two diseases, such as MS and celiac disease may share susceptibility loci but that the direction of the effect (in terms of the risk allele) is the same in only 50% of the loci. In the other 50%, the allele associated with risk in one disease is protective in the other [58]. Analyzing the functional consequences of these loci in different combinations that recognize an underlying molecular pathway architecture may help us to understand how the same variants can lead to very different diseases. In addition, such analyses will uncover key nodes downstream of several different variants that (1) may be excellent targets for drug development and (2) will need to be considered in drug development as treatment for one disease may push immune responses in a direction that precipitates a second inflammatory

disease while treating the first one. Thyroiditis as an adverse effect of alemtuzumab treatment in MS is one such example [59]. Overall, the immune system is highly integrated and even tissue-specific inflammatory diseases have immune alterations that are seen in the peripheral circulation that affect most immune cells directly or indirectly. Thus, it is premature to focus analyses on intermediate traits of a specific cell type for a given disease, and communities of investigators should instead focus on developing large-scale approaches that perform integrated analyses of multiple cell types that will inform the study of multiple different diseases. This is a realistic goal for the near future and will enable the integration of molecular signals across cell types – ideally in different activation contexts - that will yield higher order perspectives of immune pathways for more global approaches to immune modulation in inflammatory diseases.

Acknowledgments

We thank Towfique Raj and Manik Kuchroo for creating Figure 1.

References

- Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, Vermeire S, Dewit O, de Vos M, Dixon A, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 2007; 3:e58. [PubMed: 17447842]
- Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature.* 2010; 466:714–719. [PubMed: 20686566] Musunuru et al. 2010, *Nature**The authors demonstrate that common noncoding polymorphisms in the 1p13 region associated with myocardial infarction and low lipoprotein cholesterol (LDL-C) in humans alter expression of the SORT1 gene and alter lipoprotein metabolism. This is one of few examples integrating genome-wide association studies, regulatory variation, and in-depth functional studies of non-coding genetic variants, ultimately elucidating novel pathways contributing to disease.
- Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. Genetics of gene expression and its effect on disease. *Nature.* 2008; 452:423–428. [PubMed: 18344981]
- Prescott NJ, Dominy KM, Kubo M, Lewis CM, Fisher SA, Redon R, Huang N, Stranger BE, Blaszczyk K, Hudspeth B, et al. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum Mol Genet.* 2010; 19:1828–1839. [PubMed: 20106866]
- Moffatt MF, Kabisch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufer A, Rietschel E, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature.* 2007; 448:470–473. [PubMed: 17611496]
- Okada Y, Shimane K, Kochi Y, Tahira T, Suzuki A, Higasa K, Takahashi A, Horita T, Atsumi T, Ishii T, et al. A genome-wide association study identified AFF1 as a susceptibility locus for systemic lupus erythematosus in Japanese. *PLoS Genet.* 2012; 8:e1002455. [PubMed: 22291604]
- De Jager PL, Baecher-Allan C, Maier LM, Arthur AT, Ottoboni L, Barcellos L, McCauley JL, Sawcer S, Goris A, Saarela J, et al. The role of the CD58 locus in multiple sclerosis. *Proc Natl Acad Sci U S A.* 2009; 106:5264–5269. [PubMed: 19237575]
- Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science.* 2009; 325:1246–1250. [PubMed: 19644074] Dimas et al. 2009, *Science**This study performed expression quantitative trait locus (eQTL) mapping in B-cells, T-cells, and fibroblasts of a single set of individuals, revealing that the majority (70–80%) of *cis*-eQTL associations are cell-type specific. This is the first study to compare eQTLs of different cell types within a single cohort, and illustrates the importance of cell-type in elucidating the full compendium of functional regulatory variation.
- Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC. Genetics of gene expression in primary immune cells identifies cell type-

specific master regulators and roles of HLA alleles. *Nat Genet.* 2012; 44:502–510. [PubMed: 22446964] Fairfax et al 2012, *Nature Genetics**The authors characterized the regulatory landscape of both purified primary B-cells and monocytes in nearly three hundred human individuals, and identified pathways and mechanisms underlying susceptibility to human autoimmune disease, as well as *trans*-regulatory effects of alleles within the major histocompatibility complex (MHC) region, a key region for immune-mediated diseases.

10. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* 2011; 7:e1002003. [PubMed: 21304890]
11. 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:e107. [PubMed: 18462017]
12. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One.* 2010; 5:e10693. [PubMed: 20502693]
13. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* 2012; 8:e1002639. [PubMed: 22532805]
14. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, et al. Population genomics of human gene expression. *Nat Genet.* 2007; 39:1217–1224. [PubMed: 17873874] Stranger et al 2012, *PLoS Genetics**The authors characterized the *cis*-regulatory landscape of lymphoblastoid cell lines of over 700 individuals of eight diverse human populations. The study comprises the most diverse expression quantitative locus (eQTL) mapping study to date, and quantifies population-specific and shared functional regulatory variation.
15. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 2008; 4:e1000214. [PubMed: 18846210] Veyrieras et al 2008, *PLoS Genetics**This study used a novel Bayesian hierarchical model to create high-resolution map of the typical locations of sites that affect mRNA levels in *cis*-. The results suggest an important role for mRNA stability in determining steady-state mRNA levels.
16. Dimas AS, Nica AC, Montgomery SB, Stranger BE, Raj T, Buil A, Giger T, Lappalainen T, Gutierrez-Arcelus M, Consortium M, et al. Sex-biased genetic effects on gene regulation in humans. *Genome Research.* in press.
17. Idaghdour Y, Storey JD, Jadallah SJ, Gibson G. A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs. *PLoS Genet.* 2008; 4:e1000052. [PubMed: 18404217]
18. Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, Chen W, Weichenthal M, Ellinghaus E, Franke A, Cookson W, et al. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am J Hum Genet.* 2010; 87:779–789. [PubMed: 21129726]
19. Powell JE, Henders AK, McRae AF, Wright MJ, Martin NG, Dermitzakis ET, Montgomery GW, Visscher PM. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.* 2012; 22:456–466. [PubMed: 22183966]
20. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* 2011; 7:e1001317. [PubMed: 21383966]
21. Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC, et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet.* 2009; 5:e1000692. [PubMed: 19834560]
22. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet.* 2004; 75:1094–1105. [PubMed: 15514893]
23. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature.* 2010; 464:768–772. [PubMed: 20220758] Pickrell et al

2010, Nature*This study (together with Montgomery et al 2010, Nature) is the first expression quantitative trait locus (eQTL) study to utilize RNA-sequencing (RNA-seq) data, allowing characterization of functional regulatory variation at unprecedented resolution, including splicing variation, allele-specific expression, and novel exons and transcripts.

24. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*. 2010; 464:773–777. [PubMed: 20220756] Montgomery et al 2010, Nature*This study (together with Pickrell et al 2010, Nature) is the first expression quantitative trait locus (eQTL) study to utilize RNA-sequencing (RNA-seq) data, allowing characterization of functional regulatory variation at unprecedented resolution, including splicing variation, allele-specific expression, and novel exons and transcripts.
25. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adany R, Aromaa A, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet*. 2010; 42:295–302. [PubMed: 20190752]
26. Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum Mol Genet*. 2010; 19:122–134. [PubMed: 19825846]
27. Bullaughey K, Chavarria CI, Coop G, Gilad Y. Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum Mol Genet*. 2009; 18:4296–4303. [PubMed: 19671653]
28. Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, Liu W, Lin YS, Moloney C, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet*. 2011; 7:e1002078. [PubMed: 21637794]
29. Zhong H, Beaulaurier J, Lum PY, Molony C, Yang X, Macneil DJ, Weingarh DT, Zhang B, Greenawalt D, Dobrin R, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet*. 2010; 6:e1000932. [PubMed: 20463879]
30. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet*. 2010; 86:581–591. [PubMed: 20346437]
31. Mason CC, Hanson RL, Ossowski V, Bian L, Baier LJ, Krakoff J, Bogardus C. Bimodal distribution of RNA expression levels in human skeletal muscle tissue. *BMC Genomics*. 2011; 12:98. [PubMed: 21299892]
32. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, et al. A survey of genetic human cortical gene expression. *Nat Genet*. 2007; 39:1494–1499. [PubMed: 17982457]
33. Richards AL, Jones L, Moskvina V, Kirov G, Gejman PV, Levinson DF, Sanders AR, Purcell S, Visscher PM, Craddock N, et al. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol Psychiatry*. 2012; 17:193–201. [PubMed: 21339752]
34. Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS, Carrasquillo MM, Rowley CN, Nair AA, Middha S, et al. Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants. *PLoS Genet*. 2012; 8:e1002707. [PubMed: 22685416]
35. Murphy A, Chu JH, Xu M, Carey VJ, Lazarus R, Liu A, Szeffler SJ, Strunk R, Demuth K, Castro M, et al. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum Mol Genet*. 2010; 19:4745–4757. [PubMed: 20833654]
36. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet*. 2011; 7:e1002144. [PubMed: 21811411]
37. Gamazon ER, Nicolae DL, Cox NJ. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet*. 2011; 7:e1001292. [PubMed: 21304891]
38. 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:e1000888. [PubMed: 20369019] Nicolae et al 2010, PLoS Genetics*This is the first study to

demonstrate a global enrichment for expression quantitative trait loci (eQTLs) among trait-associated variants, e.g., those from genome-wide association studies.

39. Below JE, Gamazon ER, Morrison JV, Konkashbaev A, Pluzhnikov A, McKeigue PM, Parra EJ, Elbein SC, Hallman DM, Nicolae DL, et al. Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. *Diabetologia*. 2011; 54:2047–2055. [PubMed: 21647700]
40. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EA, Stokkers PC, van Bodegraven AA, Crusius JB, Hommes DW, et al. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Hum Mol Genet*. 2010; 19:3482–3488. [PubMed: 20601676]
41. Wallace C, Rotival M, Cooper JD, Rice CM, Yang JH, McNeill M, Smyth DJ, Niblett D, Cambien F, Tiret L, et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum Mol Genet*. 2012; 21:2815–2824. [PubMed: 22403184]
42. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*. 2010; 6:e1000895. [PubMed: 20369022] Nica et al 2010, *PLoS Genetics**This study provides a framework for assessing whether a trait association and an expression quantitative trait locus (eQTL) that co-localize to a single genomic region represent the same underlying causal variant.
43. Hu X, Kim H, Stahl E, Plenge R, Daly M, Raychaudhuri S. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am J Hum Genet*. 2011; 89:496–506. [PubMed: 21963258] Hu et al 2011, *Am J Hum Genetics**The authors developed a statistical approach to identify potentially pathogenic cell types in autoimmune diseases by using a gene-expression data set of 223 murine-sorted immune cells from the Immunological Genome Consortium. Such an approach can inform future functional experiments in relevant cell types for specific diseases.
44. Heng TS, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol*. 2008; 9:1091–1094. [PubMed: 18800157]
45. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004; 32:D277–280. [PubMed: 14681412]
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
47. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*. 2003; 31:334–341. [PubMed: 12520017]
48. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*. 2011; 39:D691–697. [PubMed: 21067998]
49. Lee PH, Bergen SE, Perlis RH, Sullivan PF, Sklar P, Smoller JW, Purcell SM. Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework. *Hum Hered*. 2011; 72:10–20. [PubMed: 21849790]
50. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*. 2010; 6
51. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009; 4:44–57. [PubMed: 19131956]
52. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009; 37:1–13. [PubMed: 19033363]
53. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*. 2011; 7:e1001273. [PubMed: 21249183] Rossin et al.

2011 PLoS Genetics*This study demonstrates that immune-mediated disease loci identified by genome-wide association studies harbor genes encoding proteins that physically interact to a greater extent than expected by chance. Identification of protein-protein interaction networks underlying disease allowed discovery of novel disease-associated variants.

54. Xia Z, Liu Q, Berger CT, Keenan BT, Kaliszewska A, Cheney PC, Srivastava GP, Castillo IW, De Jager PL, Alter G. A 17q12 allele is associated with altered NK cell subsets and function. *J Immunol.* 2012; 188:3315–3322. [PubMed: 22345646]
55. Dendrou CA, Plagnol V, Fung E, Yang JH, Downes K, Cooper JD, Nutland S, Coleman G, Himsworth M, Hardy M, et al. Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat Genet.* 2009; 41:1011–1015. [PubMed: 19701192]
56. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, Stranger BE, De Jager PL. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. *Am J Hum Genet.* 2012; 90:720–726. [PubMed: 22482808]
57. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* 2011; 7:e1002254. [PubMed: 21852963] Cotsapas et al 2011 PLoS Genetics*This study characterized the extent of sharing of the genetic basis of disease susceptibility among seven autoimmune and inflammatory diseases, finding evidence that 44% of known immune-mediated disease variants are associated to multiple, but not all, immune-mediated diseases.
58. Patsopoulos NA, Esposito F, Reischl J, Lehr S, Bauer D, Heubach J, Sandbrink R, Pohl C, Edan G, Kappos L, et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann Neurol.* 2011; 70:897–912. [PubMed: 22190364]
59. Coles AJ, Wing M, Smith S, Coraddu F, Greer S, Taylor C, Weetman A, Hale G, Chatterjee VK, Waldmann H, et al. Pulsed monoclonal antibody treatment and autoimmune thyroid disease in multiple sclerosis. *Lancet.* 1999; 354:1691–1695. [PubMed: 10568572]
60. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010; 26:2336–2337. [PubMed: 20634204]

Highlights

- Genome-wide association studies have identified many genetic variants associated with inflammatory disease risk.
- Gene expression may help to elucidate causal genes and functional mechanisms underlying disease susceptibility.
- Inflammatory disease variants are enriched for gene regulatory function.
- Because of cell-type and tissue specificity of gene regulation, functional studies need to be performed in relevant cell types.
- Analyses evolve into a more “systems genomic” approach where genomic data are integrated to identify causal mechanisms and pathways for disease.

Figure 1. Colocalization of *cis*-expression quantitative trait loci and Multiple Sclerosis GWAS signals in the CD40 locus

The top panel reports p-values for *cis*-regulatory associations of SNPs on CD40 RNA expression levels in peripheral blood mononuclear cells of multiple sclerosis patients; $-\log_{10}(\text{p-value})$ is reported on the y-axis. The lower panel reports the Multiple Sclerosis (MS) GWAS signal ($-\log_{10}(\text{p-value})$) for each SNP over the same chromosomal segment (1 Mb total) [58]. The RefSeq genes in the region are shown at the bottom of the figure. The LD (as quantified by r^2) for each SNP with the index MS-associated CD40 SNP (rs6074022) is illustrated with the use of colors, as indicated in the top right of the figure. The figure was generated using LocusZoom [60].

Table 1

Database or browser name	web-link	Reference
eQTL.chicago.edu browser	http://eatl.uchicago.edu/cgi-bin/gbrowse/eQTL/	Veyrieras et al 2008
SNP Express	http://compute1.lsrc.duke.edu/software/SNPExpress/1_database.php	Heinzen et al 2008
mRNAbay SNP browser	http://www.sph.umich.edu/csg/liang/asthma/	Dixon et al 2007, Moffatt et al 2007
SCAN:SNP and CNV Annotation Database	http://www.scandb.org	Gamazon et al 2010
Genevar: GENe Expression VARIation	http://www.sanger.ac.uk/software/analysis/genevar/	Yang, T.P. et al 2010
seeQTL	http://www.bios.unc.edu/research/genomic_software/seeQTL/	Xia, K. et al 2012
GTEx (Genotype-Tissue Expression) eQTL Browser	http://www.ncbi.nlm.nih.gov/gtex/test/GTEX2/gtex.cgi	
WebQTL	http://webqtl.org	Wang et al 2003