Expression Quantitative Trait Loci Analysis Identifies Associations Between Genotype and Gene Expression in Human Intestine

BOYKO KABAKCHIEV and MARK S. SILVERBERG

Zane Cohen Centre for Digestive Diseases, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada

See Covering the Cover synopsis on page 1328; see editorial on page 1339.

BACKGROUND & AIMS: Genome-wide association studies have greatly increased our understanding of intestinal disease. However, little is known about how genetic variations result in phenotypic changes. Some polymorphisms have been shown to modulate quantifiable phenotypic traits; these are called quantitative trait loci. Quantitative trait loci that affect levels of gene expression are called expression quantitative trait loci (eQTL), which can provide insight into the biological relevance of data from genome-wide association studies. We performed a comprehensive eQTL scan of intestinal tissue. **METHODS:** Total RNA was extracted from ileal biopsy specimens and genomic DNA was obtained from whole-blood samples from the same cohort of individuals. Cis- and trans-eQTL analyses were performed using a custom software pipeline for samples from 173 subjects. The analyses determined the expression levels of 19,047 unique autosomal genes listed in the US National Center for Biotechnology Information database and more than 580,000 variants from the Single Nucleotide Polymorphism database. **RESULTS:** The presence of more than 15,000 cis- and trans-eQTL was detected with statistical significance. eQTL associated with the same expression trait were in high linkage disequilibrium. Comparative analysis with previous eQTL studies showed that 30% to 40% of genes identified as eQTL in monocytes, liver tissue, lymphoblastoid cell lines, T cells, and fibroblasts are also eQTL in ileal tissue. Conversely, most of the significant eQTL have not been previously identified and could be tissue specific. These are involved in many cell functions, including division and antigen processing and presentation. Our analysis confirmed that previously published cis-eQTL are single nucleotide polymorphisms associated with inflammatory bowel disease: rs2298428/UBE2L3, rs1050152/SLC22A4, and SLC22A5. We identified many new associations between inflammatory bowel disease susceptibility loci and gene expression. **CONCLUSIONS: eQTL analysis of intestinal tissue supports findings that some eQTL remain stable across cell types, whereas others are specific to the sampled location. Our findings confirm and expand the number of known genotypes associated with expression and could help elucidate mechanisms of intestinal disease.**

Keywords: SNPdb; IBD; Transcriptomics; Systems Biology.

Over the past decade, genome-wide association studies (GWAS) have been instrumental in identifying numerous loci related to complex diseases. Since the first GWAS paper was published in 2005,[1] linking age-related macular degeneration to single nucleotide polymorphisms (SNPs) in the *CFH* gene, the field of genetic research has seen a proliferation in the application of this approach. The catalogue of published GWAS, curated by the National Institutes of Health, lists more than 1200 studies spanning more than 600 phenotypic traits. These include diverse disorders such as asthma,[2] autism,[3] rheumatoid arthritis,[4] diabetes,[5] glaucoma,[6] and Parkinson's disease[7] to name a few. Association studies, however, are not limited to categorical outcomes but can also be extended to measurable traits, such as body weight,[8] calcium levels,[9] vitamin B_{12} levels,[10] and blood pressure.[11] In these situations, quantitative trait loci analysis has been applied successfully in correlating levels of a trait of interest with genotype. Perhaps the most natural trait to be associated with variations in the genome is the immediate product of transcribed genes: messenger RNA. Indeed, such an approach combines 2 genome-wide technologies together for a systems biology treatment of physiologic problems.

The core hypothesis behind expression quantitative trait loci (eQTL) analysis is that polymorphic sites in the genome, such as SNPs, could have a tangible effect on gene regulation by altering the coding or promoter sequences of genes, their splicing junctions, or other regulatory elements. All of these regions affect the rate at which genes are transcribed, which isoforms are preferentially expressed, and how stable the final messenger RNA product is. Thus, SNPs suspected of affecting gene expression (eSNPs) can be tested with associative statistics. Two types of eQTL can generally be differentiated from one another: cis and trans. Polymorphic sites within chromo-

Abbreviations used in this paper: eQTL, expression quantitative trait loci; eSNP, expressed single nucleotide polymorphism; FDR, false discovery rate; GC, GenCall; GWAS, genome-wide association studies; IBD, inflammatory bowel disease; kb, kilobase; LD, linkage disequilibrium; MAF, minor allele frequency; mb, megabase; SNP, single nucleotide polymorphism.

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2013.03.001

somal proximity of a gene affecting its expression are considered to be cis regulators. In contrast, elements elsewhere on the genome are thought to be acting in trans. The distinction between cis and trans, however, is not always well defined. With the exception of a few straightforward possibilities when the polymorphic site falls within the start and end coordinates of a gene, or when it is on a different chromosome altogether, labeling a locus as either cis or trans is determined by a rigid distance from a gene measured in base pairs. Commonly used intervals around genes to define cis action have ranged from a few tens of thousands of bases to millions.[12–15]

A number of eQTL studies have been published to date on a variety of tissues and cells. Some of these include monocytes,[14] liver tissue,[16] and brain tissue.[17] At least 3 public databases allow for quick and easy access to the significant results reported in a few published papers. Combining these data, however, is challenging not only because different groups approach eQTL analysis with a different statistical framework, but also because experimental techniques vary considerably. In most instances, microarray technology is used to measure gene expression, but this field has also seen the advent of RNA sequencing.[18–21] Genotyping platforms by different manufacturers, or indeed across variations of the same platform, provide dissimilar coverage of genetic markers, further complicating comparison across studies. Nevertheless, at least 30% of eQTL appear to be stable between tissues and cell types, with one study estimating this number to be as high as 50% to 60%.[22] However, some eQTL do exhibit tissue specificity and it is paramount to identify these, especially in the context of disease. Furthermore, for the majority of disorders that affect a single organ or a limited number of tissues, eQTL that have a tangible effect on phenotype might be undetectable at a different anatomic location. In addition, some eQTL present in multiple tissues have been shown to exhibit completely opposite effects depending on the cell type.[23]

A comprehensive analysis of the GWAS data spanning many different published studies indicated that trait-associated loci, especially those pertaining to complex diseases, are enriched for being eQTL as well.[24] Specifically, Nicolae et al estimated that approximately 17% of Crohn's disease–associated SNPs could be eQTL in lymphoblastoid cell lines. This estimation raises the question whether this number of eQTL is similar in gastrointestinal tissue and what content is preserved. Therefore, the main goal of our study was to identify the eQTL that are active in the human ileum. In addition, using Crohn's disease as a template, we aimed to improve the current understanding of this disease by providing context for a number of the currently known susceptibility SNPs.

Materials and Methods

Subject Cohort

Individuals who underwent ileal pouch–anal anastomosis following colectomy were recruited at Mount Sinai Hospital in Toronto, Ontario, Canada, in compliance with the hospital's research ethics board. The cohort consisted of subjects with a diagnosis of ulcerative colitis or familial adenomatous polyposis, which are disorders that primarily affect the large intestine. The subjects were recruited at least 1 year after closure of their ileostomy. An extensive panel of clinical information and biospecimens were collected on recruitment, including a clinical disease activity index, physician's global assessment, complete history of medication use, endoscopic and histopathologic evaluation of the prepouch ileum, complete blood cell count, C-reactive protein level, whole blood for DNA extraction, and tissue biopsy specimens for RNA analysis. Where necessary, these data were collected in a format compatible with both the Heidelberg Pouchitis Activity Score and Pouchitis Disease Activity Index to assess inflammatory state.[25,26]

Messenger RNA Extraction and Quantification

Two tissue biopsy specimens were obtained from endoscopically and histologically normal prepouch ileum of every eligible subject. The samples were then immediately suspended in RNAlater (Qiagen, Venlo, Netherlands) stabilizing reagent to deter RNA degradation and stored at −80°C. Total RNA was extracted with the miRNeasy Mini Kit (Qiagen) in 2 batches. NanoDrop 1000 (Thermo Fisher Scientific, Waltham, MA) and Bioanalyzer 2100 (Agilent, Santa Clara, CA) were used to determine RNA concentration, quality, and purity. Only samples with an RNA integrity number $\geq$5.0 were considered for further analysis.[27] Additional information on RNA integrity number cutoffs is provided in Supplementary Materials and Methods.

A total of 400 ng of RNA from samples that passed quality control were amplified with the Ambion WT Expression Kit (Life Technologies, Carlsbad, CA). A total of 5.5 μg of complementary DNA per sample was then labeled and hybridized to Human Gene 1.0 ST arrays (Affymetrix, Santa Clara, CA) in a Fluidics Station 450 (Affymetrix) using standard protocol FS450_0007 with GeneChip WT Terminal Labeling and Controls Kit (Affymetrix) and GeneChip Hybridization, Wash, and Stain Kit (Affymetrix). GeneChip Scanner 3000 (Affymetrix) was used to scan the completed arrays. Summarized probe cell intensity data were generated with an Affymetrix GeneChip Command Console. Finally, probe-level summarization files were produced and the data were background adjusted, normalized, and log transformed with the robust multiarray average algorithm in Affymetrix Expression Console.[28]

The empirical Bayes method described by Johnson et al[29] was applied to the normalized data to correct for batch effects that may have resulted from a nonlinear sample extraction and microarray processing schedule. Lastly, duplicate and ambiguous Affymetrix probe sets (Release 32) as well as those that no longer mapped to a gene in the current human genome build (GRCh37.p5) were removed from further analysis.

Genotyping

Parallel to gene expression quantification, genomic DNA was extracted from whole-blood samples collected at the time of recruitment. The Gentra Puregene Blood Kit (Qiagen) was used to separate and lyse white blood cells for purification of DNA from blood biospecimens. Extracted DNA elutes were normalized at 50 ng/μL and further analyzed in a 96-well plate format. Samples were hybridized to either HumanOmniExpress or HumanOmni2.5 BeadChips (Illumina, San Diego, CA), and the iScan system (Illumina, San Diego, CA) was used to scan the