NIH Public Access
Author Manuscript
Gastroenterology. Author manuscript; available in PMC 2014 June 01.

Published in final edited form as:
Gastroenterology. 2013 June ; 144(7): 1488–1496.e3. doi:10.1053/j.gastro.2013.03.001.

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

Abstract

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

© 2013 by the AGA Institute

Reprint requests Address requests for reprints to: Boyko Kabakchiev, 60 Murray Street, Room L3-004, Toronto, Ontario M5T 3L9, Canada. kabakchiev@lunenfeld.ca; fax: (416) 586-5932; or Mark Silverberg, 600 University Avenue, Room 441, Toronto, Ontario M5G 1×5, Canada. msilverberg@mtsinai.on.ca; fax: (416) 619-5524..

Web Resources DAVID Functional Annotation Bioinformatics Microarray Analysis: http://david.abcc.ncifcrf.gov/. The GEO accession number for the gene expression microarray and genotyping data reported in this paper is GSE40292.

Supplementary Material Note: To access the supplementary material accompanying this article, visit the online version of Gastroenterology at www.gastrojournal.org, and at http://dx.doi.org/10.1053/j.gastro.2013.03.001.

Conflicts of interest The authors disclose no conflicts.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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-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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 ≥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]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 arrays. Genotypes were then called in GenomeStudio (Illumina). Samples genotyped on the larger platform, HumanOmni2.5, were scaled back to the overlap with HumanOmniExpress coverage. Genotypes with GenCall (GC) scores <0.2 and samples with call rates <95% were excluded as failed markers and samples, respectively. Only SNPs with call rates >95%, minor allele frequency (MAF) >5%, and Hardy–Weinberg equilibrium χ^2 P values greater than 10^{-6} were considered. Raw genotyping data are available on request.

Bioinformatics

A custom standalone software application, eQTLA, was created to address the analytical issues surrounding eQTL studies (Supplementary Figure 1). eQTLA is freely available as precompiled executables as well as source code in the Supplementary Material. The core motivation behind the design of this software was to create a tool that can efficiently perform eQTL analysis on a single workstation with limited hardware resources.

eQTLA is capable of performing both cis- and trans-eQTL analysis given a set of raw genetic marker data and expression values. Currently, a C++ implementation of the nonparametric Kruskal–Wallis one-way analysis of variance[30] is the test of significance applied to genotype comparison, forgoing the need for normally distributed data. Nevertheless, the addition of other statistical tests to the framework of eQTLA is straightforward due to its object-oriented nature. Raw P values are adjusted for multiple testing by either Bonferroni or Benjamini and Hochberg false discovery rate (FDR) correction.[31]

An optional feature in eQTLA can identify markers in close proximity that are highly correlated due to apparent linkage disequilibrium (LD). This is accomplished by pairwise calculation of either D' or r^2 values between markers within a given genomic interval. A C++ implementation of Gregory Warnes's *LD* function from the R package *genetics* is used for this operation. Each group of markers in high LD is assigned a unique locus ID. Post-hoc adjustment can be applied to eQTL data by amalgamating markers from the same locus and assigning the median $\log_{10}$ converted P value to the cluster.

eQTL and Statistical Analysis

Information on the genomic coordinates of genes based on human genome build GRCh37.p5 was downloaded from the National Center for Biotechnology Information. Cis-eQTL analysis was centered about regions of DNA, dubbed "windows," containing an autosomal gene of interest along with the 50-kilobase (kb) sequence both upstream and downstream from its starting and end points. Alternative analysis with 1-megabase (mb) windows was also performed to encompass the wide range of window sizes used in previously reported

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

eQTL studies. Only SNPs located in these intervals were included in this cis analysis, but those located in regions where intervals overlap were used more than once. Results were generated for individual markers as well as for clusters of markers in high LD. SNPs on the same chromosome, no farther than 200 kb apart for the 50-kb window analysis and 1.1 mb for the 1-mb window analysis, and with r^2 values ≥0.5 were considered to be in high LD. Raw P values were corrected for multiple testing by the FDR method at an α level of 5%.

Two conservative variations of trans-eQTL analysis were also performed. One approach involved the use of human interactome data. Specifically, information from the Human Protein Reference Database and Biological General Repository for Interaction Datasets was combined in a single set of binary proteinprotein interactions. Associative analysis was performed between SNPs located in the genomic regions of genes and the expression measurements of their interacting partners. The underlying hypothesis of this inquiry was that polymorphic sites in genes may result in altered protein structure, protein stability, or isoform ratio and thereby have no immediate effect on the expression of those genes but rather on their downstream pathway interactors. Raw P values were corrected for multiple testing by the FDR method.

A secondary method for trans-based analysis built on the initial cis-eQTL results. Genotypes that were significantly associated with gene expression in cis at a P value cutoff of .05 after Bonferroni correction were tested in trans. This was performed for all expression measurements; however, an exclusion window of the same size as defined in the cis analysis was set about every gene to avoid duplicate associations.

Comparison With Other Tissues

Significant cis-eQTL results from this study were compared with other previously reported cis-eQTL data. These included the following tissue sources: monocytes,[14] liver,[16] lymphoblastoid cell lines,[32] T cells,[33] and fibroblasts.[33] The comparative analysis focused on genes in eQTL pairs rather than eSNPs to reduce bias from different genotyping platforms used in these studies. In addition, only autosomal genes with expression that was successfully measured in ileal tissue using the Affymetrix platform were considered from the other sources. Variations of the comparative analysis based on 50-kb and 1-mb windows as well as an α level of 5% and 10% after FDR correction for multiple testing were all performed.

eQTL and Inflammatory Bowel Disease

Significant cis-acting eSNPs in this study were examined for possible replication of a small number of eQTL that have been linked to inflammatory bowel disease (IBD) in the past. Additionally, 163 genome-wide associated IBD SNPs established using the ImmunoChip platform[34] were evaluated in a subset of subjects for which ImmunoChip data were available. These SNPs were selected for a targeted, hypothesis-driven eQTL analysis performed both in cis and in trans fashion.

Results

Basic Measures of Data Quality

Following quality control, 19,908 unambiguous Affymetrix Human Gene 1.0 ST array probe sets remained in the expression data set, of which 19,047 mapped to autosomal genes. Genotyping was completed successfully on 173 individuals. The average observed call rate per sample was 97.3%, and the minimum was 97.2%. Similarly, the average call rate per SNP was 97.3%, with 97.4% of SNPs having call rates ≥95%. Other quality control measures such as the Illumina GC score were also high. The mean GC score for all SNPs on

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

the HumanOmniExpress platform was 0.84, and 82.9% of the SNPs had a score ≥0.8 (Supplementary Figure 2*A*). Respectively, the mean GC score for all SNPs on the HumanOmni2.5 BeadChips was 0.81, with 70% of SNPs having scores ≥0.8 (Supplementary Figure 2*B*). In total, considering all samples in the study, there were 581,633 SNPs with a call rate >95%, GC score >0.2, MAF >5%, and Hardy–Weinberg equilibrium χ^2 *P* value $>10^{-6}$.

Of the 173 subjects, 153 (88.4%) described themselves as white, 18 (10.4%) as Asian, and 2 (1.2%) as black.

eQTL Analysis

Cis-eQTL analysis based on these quality-trimmed data and focusing on SNPs located within 50-kb windows around each gene detected the presence of 15,091 statistically significant cis-eQTL associated with 2629 genes (Supplementary Table 1). Twelve of the most significant eQTL are depicted in Figure 1. The MAFs of significant eSNPs were spread widely between 5% and 50%, but the distribution was skewed toward the higher frequencies (Figure 2). The expression traits pertain to many aspects of cellular function from division to antigen processing and presentation. For instance, endoplasmic reticulum aminopeptidase 2 (*ERAP2*), with the function of trimming antigen peptides in the endoplasmic reticulum for presentation on major histocompatibility complex I, was the strongest expression trait detected in this study.[35] Immediately following *ERAP2* in significance was X-ray radiation resistance associated 1 (*XRRA1*). Up-regulated expression of *XRRA1* has been linked to reduced capacity for DNA damage repair in human cell lines.[36] Overall, Gene Ontology classification based on the molecular function of these 2629 genes indicated that the set was enriched for oxidoreductase and dehydrogenase activity but also for KRAB domain containing zinc finger transcription factors (Table 1). The data were very similar when 1-mb windows were used instead. This approach resulted in 14,535 significant cis-eQTL associated with a total of 1811 genes (Supplementary Table 1). Of these, 1520 (83.9%) were significant expression traits at the 50-kb window levels as well.

Cis-eQTL associated with the same expression trait were in high LD (Figure 3). Significant eSNPs within 50-kb windows were grouped into clusters based on the inferred r^2 values between them. The median size of these clusters was 2; however, in some instances this number was as high as 33. Sixty-five percent of all cis-eQTL clusters contained 2 or more eSNPs (Figure 4). The total number of significant clusters was 4196 (Supplementary Table 2) and all of the associated expression traits were present in the single SNP–based results, although the reverse statement was not true. Notably, the concordance of significant expression traits between 1-mb and 50-kb window analyses increased to 92.7% after correction for high LD.

Trans-eQTL analysis resulted in far fewer significant associations. Using human interactome data, only 4 eSNPs survived correction for multiple testing. Furthermore, all of these eQTL had already been identified as cis acting and their presence in trans was an artifact of the genomic proximity between the respective interacting genes. Trans analysis of 50-kb window-based cis-eQTL proved to be more successful (Supplementary Table 3). A total of 291 of these also showed evidence of trans effects. Although the majority of these eSNPs were located on the same chromosome as the gene and possibly represent cis effects that extend beyond 50 kb, 9 trans-eQTL spanned different chromosomes.

Comparison With Other Tissues

Comparative analysis with other eQTL studies showed that a substantial fraction of eQTL in the ileum were also eQTL in other tissues (Figure 5). A total of 30.2%, 36.9%, 30.4%,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

29.3%, and 32.8% of the significant expression traits reported in monocytes, hepatic tissue, lymphoblastoid cell lines, T cells, and fibroblasts, respectively, were also significant in ileal tissue at an α level of 5% and using 50-kb windows (Supplementary Table 4). If instead an α level of 10% was applied to the data as suggested by Zeller et al,[14] the overlap between these tissues and the intestine increased by 6.5% on average. Increasing the window size to 1 MB had a detrimental effect on the observed overlap between cell types.

eQTL and IBD

Finally, in a case study applying eQTL data to a specific disease, expression traits and eSNPs pertinent to IBD were identified. In a first-step analysis, previously reported associations rs2298428/UBE2L3 and rs1050152/SLC22A4 were confirmed and the novel eQTL rs1050152/SLC22A5 was also identified. Conversely, the previously reported eSNP, rs2927488/BCL3, was not replicated and rs2631372/SLC22A5 could not be confirmed because the SNP was not measured in this study. An additional analysis encompassing 142 subjects for which ImmunoChip data were available was performed to evaluate 155 IBD-associated SNPs that passed the same quality control measures as the OmniExpress data. Of these 155 SNPs, 24 (15%) exhibited cis effects in 27 eQTL pairs at an α level of 5% and windows size of 50 kb (Table 2). Alternatively, using windows of size 1 mb and an α level of 10%, the number of significant cis-acting eSNPs was also 24 (15%) (Supplementary Table 5). After correction for multiple testing, 2 eQTL reached significance in trans.

Discussion

This study in human small intestinal tissue strongly corroborates the hypothesis of overarching gene expression changes associated with genotype. The vast majority of observed eQTL were detected in cis; however, the precise number was dependent on the choice of window size. Interestingly, enlarging the cis window to 1 mb from 50 kb did not result in substantial inflation of detected cis-eQTL. On the contrary, due to lower statistical power given the increased number of multiple comparisons combined with expansion of clusters in high LD, there was vast reduction in the number of independent eQTL that could be detected. Further substantiating the reduced diversity using larger window sizes was the comparative analysis of cis-eQTL with other tissues. At 1 mb, fewer expression traits overlap across different cell types.

However, interpretation of eQTL equivalence between tissues must be considered in the context of heterogeneity of cell types. Ileal tissue, just like liver tissue, is not homogeneous in terms of cell type composition; therefore, eQTL similarities are not necessarily equally applicable to each subset of cells. In fact, cellular differences are expected to result in changes in eQTL at different anatomic locations. Conversely, overlap of eQTL between fibroblasts and ileal tissue, for instance, is at least fractionally due to the presence of fibroblasts in the human ileum.

Identification of SNPs in high apparent LD and statistical retesting with these newly defined clusters did not qualitatively alter the detected cis-eQTL. The benefit of completing this analysis was in reducing undue emphasis on highly correlated genotypes in the data set, thereby more closely following the assumption of independent testing and eliminating weak associations based on a single SNP. Significant expression traits were more highly concordant between different window sizes. In addition, the results were less biologically misleading compared with single SNP-based data. It is likely that, for most significant eQTL, the exact causative polymorphic site was not measured by the genotyping platform and its effect was detected only through a nearby marker in high LD.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Trans-eQTL analysis was also very sensitive to statistical power. As other investigators have reported previously, trans effects tend to be weaker and replicate less often.[14,16] From a biological standpoint, given the large genomic distance between causal locus and affected gene, trans-eQTL more than likely influence expression through an intermediary such as a transcription factor or a microRNA regulator. Under these circumstances, any other factors that partake in this regulatory process could dilute the initial signal by adding noise. For instance, trans-eQTL analysis of IBD-associated SNPs resulted in many associations with P values as low as 10^{-5} and 10^{-6}; however, none of these survived correction for more than 1.4 million total comparisons given the sample size limitations of this study.

To contrast these data, cis-eQTL analysis of IBD susceptibility loci identified the expression of a number of genes as likely qualitative traits resulting from these polymorphic sites. In the past, genes in the genomic vicinity of IBD-linked polymorphisms have been prioritized as potential factors in disease mechanism, but with the exception of *FUT2*,[37,38] *MMEL1*,[39] *UBE2L3*,[40] *SLC22A5*,[41–43] *HLA-DQA1*,[44] and *DAP*,[39] all of the remaining genes (Table 2) in eQTL pairs are novel direct biological associations with IBD. To completely confirm these findings, a replication in an independent cohort will be necessary. However, these data represent a case study of how the mechanistic relevance of SNPs from GWAS can be refined successfully by eQTL analysis.

Conclusions

eQTL analysis of intestinal tissue substantiates reports in the literature that some eQTL remain stable across cell types while others are specific to the sampled location. Our findings not only confirm but also significantly expand the number of known genotypes associated with expression and could help elucidate the mechanisms of intestinal diseases including IBD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the following individuals for their significant contributions: members of the Silverberg Laboratory, including Joanne Stempak, Andrea Tyler, Raquel Milgrom, Lucy Zhang, Tsedey Legesse, Rachel Caplan, Matti Waterman, and Smita Halder; the surgeons and pathologists at Mount Sinai Hospital, including Dr Robin McLeod, Dr Zane Cohen, Dr Helen MacRae, and Dr Richard Kirsch; and the National Institute of Diabetes and Digestive and Kidney Diseases IBD Genetics Consortium, including Dr Judy Cho, Dr Steven Brant, Dr Richard Duerr, Dr Dermot McGovern, Dr John Rioux, and Dr Mark Silverberg.

Dr Mark Silverberg is partially supported by the Gale and Graham Wright Research Chair in Digestive Disease.

Funding Supported by the Crohn's and Colitis Foundation of Canada and the National Institute of Diabetes and Digestive and Kidney Diseases IBD Genetics Consortium (grants DK062429 and DK062423).

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

IBD	inflammatory bowel disease
kb	kilobase
LD	linkage disequilibrium
MAF	minor allele frequency
mb	megabase
SNP	single nucleotide polymorphism

References

1. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]
2. Moffatt MF, Kabesch M, Liang L, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448:470–473. [PubMed: 17611496]
3. Wang K, Zhang H, Ma D, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009; 459:528–533. [PubMed: 19404256]
4. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
5. Meigs JB, Manning AK, Fox CS, et al. Genome-wide association with diabetes-related traits in the Framingham Heart Study. BMC Med Genet. 2007; 8(Suppl 1):S16. [PubMed: 17903298]
6. Meguro A, Inoko H, Ota M, et al. Genome-wide association study of normal tension glaucoma: common variants in SRBD1 and ELOVL5 contribute to disease susceptibility. Ophthalmology. 2010; 117:1331–1338. e5. [PubMed: 20363506]
7. Maraganore DM, de Andrade M, Lesnick TG, et al. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet. 2005; 77:685–693. [PubMed: 16252231]
8. Thorleifsson G, Walters GB, Gudbjartsson DF, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet. 2009; 41:18–24. [PubMed: 19079260]
9. O'Seaghdha CM, Yang Q, Glazer NL, et al. Common variants in the calcium-sensing receptor gene are associated with total serum calcium levels. Hum Mol Genet. 2010; 19:4296–4303. [PubMed: 20705733]
10. Hazra A, Kraft P, Selhub J, et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat Genet. 2008; 40:1160–1162. [PubMed: 18776911]
11. Levy D, Larson MG, Benjamin EJ, et al. Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. BMC Med Genet. 2007; 8(Suppl 1):S3. [PubMed: 17903302]
12. Qiu W, Cho MH, Riley JH, et al. Genetics of sputum gene expression in chronic obstructive pulmonary disease. PLoS One. 2011; 6:e24395. [PubMed: 21949713]
13. Murphy A, Chu JH, Xu 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]
14. Zeller T, Wild P, Szymczak S, et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010; 5:e10693. [PubMed: 20502693]
15. Deutsch S, Lyle R, Dermitzakis ET, et al. Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. Hum Mol Genet. 2005; 14:3741–3749. [PubMed: 16251198]
16. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]
17. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010; 6:e1000952. [PubMed: 20485568]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

18. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
19. Babak T, Garrett-Engele P, Armour CD, et al. Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. BMC Genomics. 2010; 11:473. [PubMed: 20707912]
20. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011; 27:72–79. [PubMed: 21122937]
21. Lalonde E, Ha KC, Wang Z, et al. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Res. 2011; 21:545–554. [PubMed: 21173033]
22. Nica AC, Parts L, Glass D, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011; 7:e1002003. [PubMed: 21304890]
23. Fu J, Wolfs MG, Deelen P, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 2012; 8:e1002431. [PubMed: 22275870]
24. Nicolae DL, Gamazon E, Zhang W, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]
25. Moskowitz RL, Shepherd NA, Nicholls RJ. An assessment of inflammation in the reservoir after restorative proctocolectomy with ileoanal ileal reservoir. Int J Colorectal Dis. 1986; 1:167–174. [PubMed: 3039030]
26. Sandborn WJ, Tremaine WJ, Batts KP, et al. Pouchitis after ileal pouch-anal anastomosis: a Pouchitis Disease Activity Index. Mayo Clin Proc. 1994; 69:409–415. [PubMed: 8170189]
27. Schroeder A, Mueller O, Stocker S, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006; 7:3. [PubMed: 16448564]
28. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003; 4:249–264. [PubMed: 12925520]
29. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127. [PubMed: 16632515]
30. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952; 47:583–621.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B (Methodological). 1995; 57:289–300.
32. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. Nat Genet. 2007; 39:1217–1224. [PubMed: 17873874]
33. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009; 325:1246–1250. [PubMed: 19644074]
34. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–124. [PubMed: 23128233]
35. Tanioka T, Hattori A, Masuda S, et al. Human leukocyte-derived arginine aminopeptidase. The third member of the oxytocinase sub-family of aminopeptidases. J Biol Chem. 2003; 278:32275–32283. [PubMed: 12799365]
36. Mesak FM, Osada N, Hashimoto K, et al. Molecular cloning, genomic characterization and over-expression of a novel gene, XRRA1, identified from human colorectal cancer cell HCT116Clone2_XRR and macaque testis. BMC Genomics. 2003; 4:32. [PubMed: 12908878]
37. McGovern DP, Jones MR, Taylor KD, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet. 2010; 19:3468–3476. [PubMed: 20570966]
38. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta- analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–1125. [PubMed: 21102463]
39. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011; 43:246–252. [PubMed: 21297633]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

40. Fransen K, Visschedijk MC, van Sommeren S, 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. Peltekova VD, Wintle RF, Rubin LA, et al. Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nat Genet. 2004; 36:471–475. [PubMed: 15107849]

42. Noble CL, Nimmo ER, Drummond H, et al. The contribution of OCTN1/2 variants within the IBD5 locus to disease susceptibility and severity in Crohn's disease. Gastroenterology. 2005; 129:1854–1864. [PubMed: 16344054]

43. Vermeire S, Pierik M, Hlavaty T, et al. Association of organic cation transporter risk haplotype with perianal penetrating Crohn's disease but not with susceptibility to IBD. Gastroenterology. 2005; 129:1845–1853. [PubMed: 16344053]

44. Silverberg MS, Cho JH, Rioux JD, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat Genet. 2009; 41:216–220. [PubMed: 19122664]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

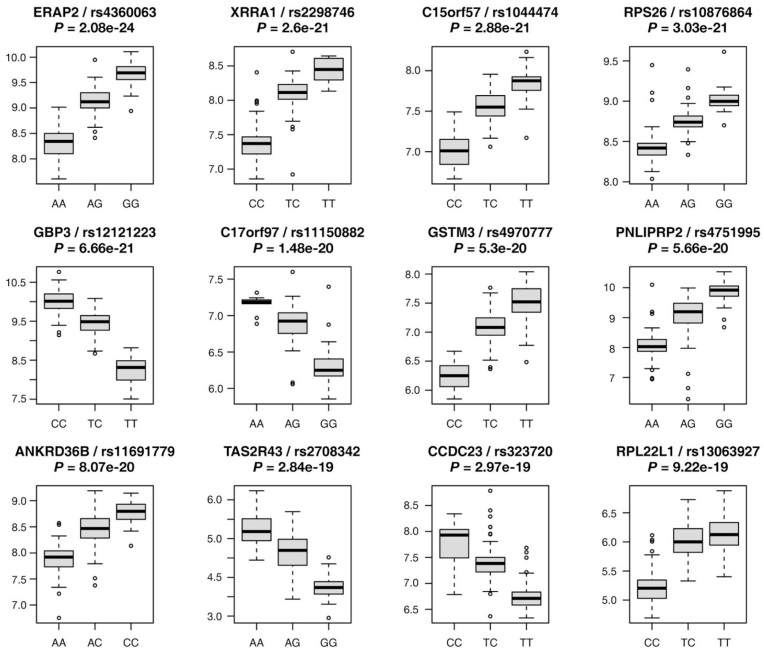

Figure 1.

Twelve significant cis-eQTL. Box plot depiction of the 12 most significant, unique cis-eQTL using a 50-kb window. The x-axis of each plot corresponds to the 3 observed SNP genotypes in forward orientation, and the y-axis represents $\log_2$-normalized gene expression values. Listed *P* values are FDR corrected.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

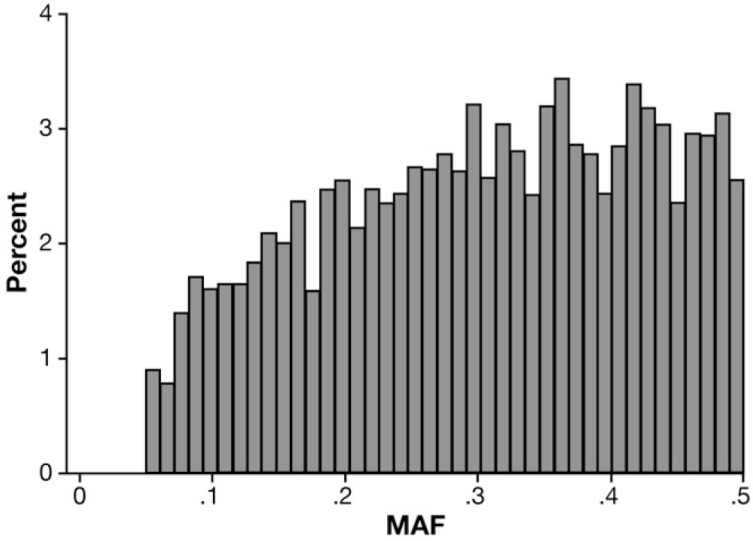

Figure 2.
MAF distribution of eSNPs. Distribution of MAF among 13,907 significant eSNPs after cis-eQTL analysis with 50-kb windows and an α level of 5%.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

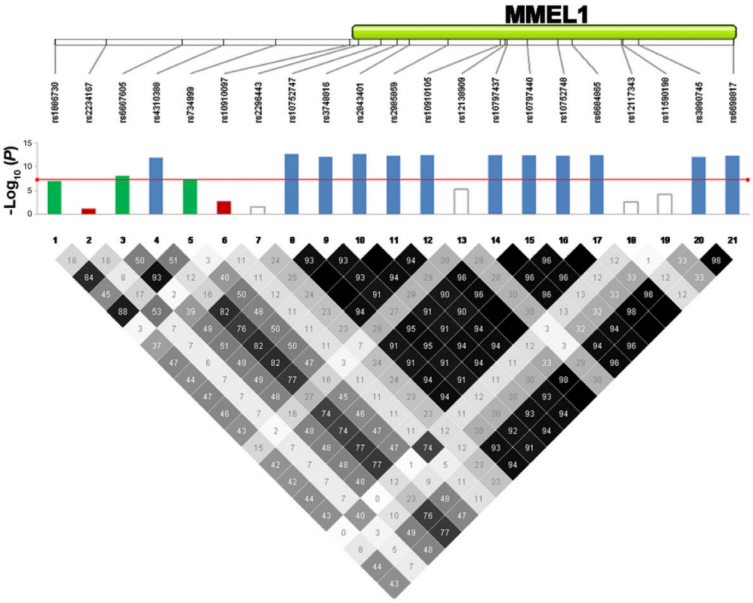

Figure 3.

Visualization of LD between SNPs surrounding the gene *MMEL1*. Visual representation of pairwise LD between markers about the *MMEL1* gene. Color intensity and values in the triangular matrix indicate the corresponding r^2 values. Bar plot heights represent the intensity of each $-\log_{10}$-transformed *P* value, and markers identified as an LD cluster by eQTLA are colored uniformly. *White bars* did not cluster with other markers. The *horizontal red line* corresponds to the Bonferroni cutoff for significance.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

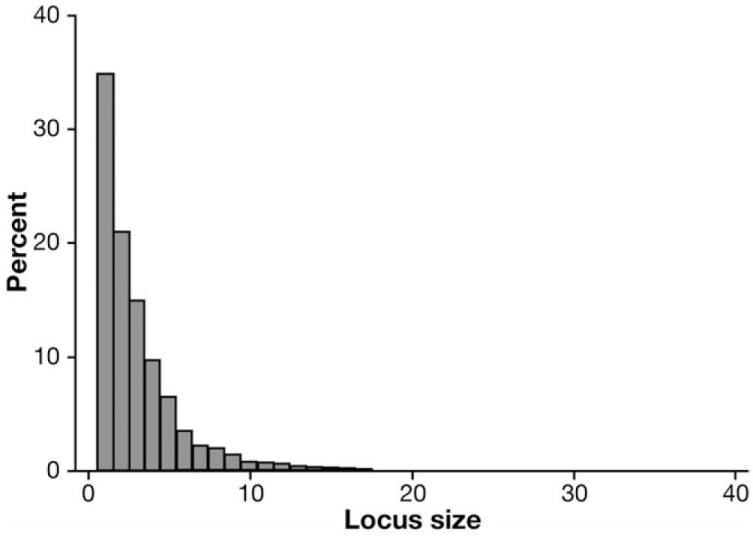

Figure 4.

Locus size distribution after LD correction with eQTLA. Distribution of locus sizes for 4196 significant clusters after cis-eQTL analysis with 50-kb windows, an α level of 5%, and subsequent LD correction.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

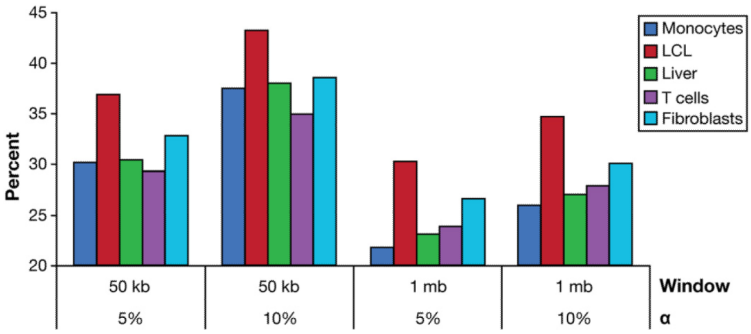

Figure 5.

eQTL comparison of human ileum with other tissues. Percentage overlap between significant cis expression traits detected in other tissues (monocytes, lymphoblastoid cell lines, liver, T cells, and fibroblasts) and human small intestinal tissue. Results are presented with varying windows sizes and α levels for the analysis in intestinal tissue.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Enriched Gene Ontology Categories Based on Molecular Function

Term	Gene count	Percent of total	*P* value	Benjamini *P* value
Oxidoreductase	122	4.7	1.90E-09	4.30E-07
Dehydrogenase	53	2.0	1.10E-06	1.20E-04
KRAB Box Transcription Factor	106	4.1	2.60E-06	2.00E-04
Transferase	141	5.4	1.00E-04	5.90E-03
Hydrolase	120	4.6	1.80E-04	8.00E-03
Reductase	38	1.5	2.40E-04	9.10E-03
Oxidase	19	0.7	1.00E-03	3.20E-02
Zinc Finger Transcription Factor	136	5.2	1.40E-03	3.80E-02

NOTE. Functional categories with significant enrichment for expression traits resulting from cis-eQTL analysis based on 50-kb windows and an α level of 5%.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Significant eQTL Linked to IBD Susceptibility

Gene	Gene chromosome	SNP	SNP chromosome	SNP position	IBD association	SNP locus	*P* value	FDR *P* value
ERAP2	5	rs1363907	5	96252803	IBD	49	8.97E-24	2.81E-21
FUT2	19	rs516246	19	49206172	CD	147	3.38E-14	5.29E-12
CTSW	11	rs568617	11	65653242	IBD	105	5.21E-13	5.44E-11
ADCY3	2	rs13407913	2	25097644	IBD	20	1.94E-10	1.52E-08
SNX32	11	rs568617	11	65653242	IBD	105	1.57E-09	9.82E-08
MMEL1	1	rs6667605	1	2502780	UC	2	5.84E-09	3.04E-07
ZNF300P1	5	rs11741861	5	150277909	IBD	54	2.37E-08	1.06E-06
FADS2	11	rs174537	11	61552680	IBD	103	5.05E-07	1.97E-05
C5orf56	5	rs2188962	5	131770805	IBD	51	1.23E-06	4.29E-05
COMMD7	20	rs6087990	20	31349908	IBD	150	1.58E-06	4.95E-05
RIT1	1	rs670523	1	155878732	IBD	12	9.75E-06	2.78E-04
SFMBT1	3	rs9847710	3	53062661	UC	38	1.41E-05	3.68E-04
CCDC122	13	rs3764147	13	44457925	CD	117	3.35E-05	8.05E-04
C13orf31	13	rs3764147	13	44457925	CD	117	7.20E-05	1.61E-03
UBE2L3	22	rs2266959	22	21922904	IBD	161	8.17E-05	1.70E-03
SLC22A5	5	rs2188962	5	131770805	IBD	51	1.29E-04	2.52E-03
HLA-DQA1	6	rs477515	6	32569691	UC	62	1.51E-04	2.78E-03
RASGRP1	15	rs16967103	15	38899190	CD	122	2.33E-04	4.04E-03
ZFP90	16	rs1728785	16	68591230	UC	131	3.33E-04	5.49E-03
DAP	5	rs2930047	5	10695526	IBD	45	5.03E-04	7.88E-03
SDCCAG3	9	rs10781499	9	139266405	IBD	89	5.78E-04	8.62E-03
FAM55A	11	rs561722	11	114386830	UC	109	6.69E-04	9.52E-03
CPSF3L	1	rs12103	1	1247494	IBD	1	1.03E-03	1.41E-02
PNKD	2	rs2382817	2	219151218	IBD	32	1.81E-03	2.36E-02
UBAC2	13	rs3742130	13	99907341	IBD	118	2.52E-03	3.15E-02
IFNGR2	21	rs2284553	21	34776695	CD	158	2.55E-03	3.08E-02
RGS14	5	rs4976646	5	176788570	IBD	57	3.95E-03	4.57E-02

NOTE. List of 27 significant cis-eQTL at an α level of 5% and windows size of 50 kb following a targeted analysis of 155 IBD-associated SNPs. These include ulcerative colitis only (UC), Crohn's disease only (CD), and both UC and CD (IBD)-associated SNPs.