nature genetics

Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease

Urko M Marigorta[1], Lee A Denson[2], Jeffrey S Hyams[3], Kajari Mondal[4], Jarod Prince[4], Thomas D Walters[5], Anne Griffiths[5], Joshua D Noe[6], Wallace V Crandall[7], Joel R Rosh[8], David R Mack[9], Richard Kellermayer[10], Melvin B Heyman[11], Susan S Baker[12], Michael C Stephens[13], Robert N Baldassano[14], James F Markowitz[15], Mi-Ok Kim[16], Marla C Dubinsky[17], Judy Cho[17], Bruce J Aronow[18], Subra Kugathasan[4,19] & Greg Gibson[1,19]

Gene expression profiling can be used to uncover the mechanisms by which loci identified through genome-wide association studies (GWAS) contribute to pathology[1,2]. Given that most GWAS hits are in putative regulatory regions and transcript abundance is physiologically closer to the phenotype of interest[2], we hypothesized that summation of risk-allele-associated gene expression, namely a transcriptional risk score (TRS), should provide accurate estimates of disease risk. We integrate summary-level GWAS and expression quantitative trait locus (eQTL) data with RNA-seq data from the RISK study, an inception cohort of pediatric Crohn's disease[3,4]. We show that TRSs based on genes regulated by variants linked to inflammatory bowel disease (IBD) not only outperform genetic risk scores (GRSs) in distinguishing Crohn's disease from healthy samples, but also serve to identify patients who in time will progress to complicated disease. Our dissection of eQTL effects may be used to distinguish genes whose association with disease is through promotion versus protection, thereby linking statistical association to biological mechanism. The TRS approach constitutes a potential strategy for personalized medicine that enhances inference from static genotypic risk assessment.

GWAS have been very successful in identifying thousands of genetic variants associated with disease, but the predictive performance of GRSs is limited by the amount of heritability they explain, which is usually low[5–8]. Given that the majority of variants discovered by GWAS are likely to influence gene regulation, risk scores based on gene expression could constitute an alternative to classical GRSs. We explored the performance of the TRS in the RISK study, which was designed to identify factors that increase risk of a complicated course of disease and included ileal biopsies from 215 patients with complication-free Crohn's disease and 35 controls profiled at diagnosis with RNA-seq[3,4,9]. After careful monitoring for 3 years, 27 of the patients with Crohn's disease progressed to stricturing or penetrating disease, allowing us to ask whether genomic profiles could be used to inform mechanisms of pathogenesis and predict disease status.

We started by considering 232 independent SNPs associated with IBD or one of its main forms—Crohn's disease or ulcerative colitis[10]. Assigning relevant genes at GWAS loci can be challenging, but eQTL studies provide an effective way to uncover which gene is likely to account for the discovered pathogenic effects. We queried the Blood eQTL browser (see URLs), a large meta-analysis of eQTL effects in peripheral blood[11], to ascertain genes regulated by IBD-predisposing variants. Around half (n = 122; 52.6%) of IBD-associated SNPs acted as or were in strong linkage disequilibrium (LD; $r^2 > 0.8$) with at least one *cis*-eQTL in peripheral blood, for a total of 157 independent candidate genes (~1.3 candidate gene per SNP; **Supplementary Table 1**).

The RNA-seq samples from the RISK study consisted of ileal biopsies, so we next asked whether the aforementioned eQTLs are also active in small intestine (Online Methods). In line with previous studies[12–14], we observed considerable sharing of signals between the two tissue types (**Supplementary Table 2**), with strong concordance in the direction of effects (70%; $P = 1.7 \times 10^{-6}$, sign test) and including just two cases with reversal of sign between blood and ileum confirmed in

[1]Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, Georgia, USA. [2]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [3]Division of Digestive Diseases, Hepatology and Nutrition, Connecticut Children's Medical Center, Hartford, Connecticut, USA. [4]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA. [5]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [6]Department of Pediatric Gastroenterology, Hepatology and Nutrition, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [7]Department of Pediatric Gastroenterology, Nationwide Children's Hospital, Ohio State University College of Medicine, Columbus, Ohio, USA. [8]Department of Pediatrics, Goryeb Children's Hospital, Morristown, New Jersey, USA. [9]Department of Pediatrics, Children's Hospital of Eastern Ontario IBD Centre and University of Ottawa, Ottawa, Ontario, Canada. [10]Section of Pediatric Gastroenterology, Baylor College of Medicine, Texas Children's Hospital, Houston, Texas, USA. [11]Department of Pediatrics, University of California, San Francisco, San Francisco, California, USA. [12]Department of Digestive Diseases and Nutrition Center, University at Buffalo, Buffalo, New York, USA. [13]Department of Pediatric Gastroenterology, Mayo Clinic, Rochester, Minnesota, USA. [14]Department of Pediatrics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [15]Department of Pediatrics, Northwell Health, New York, New York, USA. [16]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [17]Department of Pediatrics, Mount Sinai Hospital, New York, New York, USA. [18]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [19]These authors jointly directed this work. Correspondence should be addressed to G.G. (greg.gibson@biology.gatech.edu).

Received 14 January; accepted 21 July; published online 14 August 2017; doi:10.1038/ng.3936

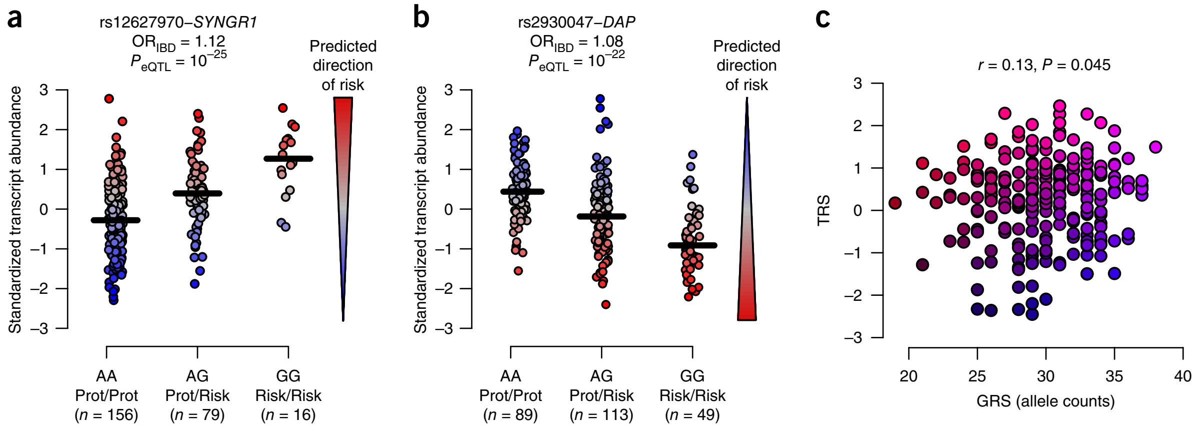

Figure 1 Transcriptional risk scores integrate GWAS and eQTL results to measure individual risk of disease based on transcript abundance. (**a**) The rs12627970[G] allele increases susceptibility to IBD and is associated with elevated expression of *SYNGR1*. Some individuals with the risk genotype (GG) show average or even low expression levels, and some individuals with the protective genotype (AA) have high expression, suggesting that abundance of *SYNGR1* provides a different estimate of individual risk of disease than genotype. Black horizontal bars denote the median expression for each genotype. OR, odds ratio. (**b**) By contrast, the rs2930047[G] risk allele is associated with lower expression of *DAP*, implying that reduced levels of *DAP* increase risk of IBD and, hence, that inversion of the *z*-score measures polarized risk of disease. (**c**) Summation of polarized transcriptional activity according to eQTL activity (left *y* axis in **a** and **b**) summed over all genes, and further standardized, is correlated with an allelic-sum GRS plotted on the *x* axis but provides an independent predictor of IBD. The *P* values in **a** and **b** correspond to the eQTL study in RISK samples, and the *P* value in **c** corresponds to a Spearman's rank correlation test.

the Genotype-Tissue Expression (GTEx) data set (*PNKD* and *RGS14*; **Supplementary Fig. 1**). This overlap indicates that eQTL effects at IBD-associated SNPs can be used to polarize gene expression relative to risk as a means to understand which allele is associated with pathogenesis at each gene. For instance, the G risk allele for IBD at rs12627970 increased abundance of *SYNGR1* (**Fig. 1a**), whereas the G risk allele at rs2930047 downregulated *DAP* (**Fig. 1b**). We can hence polarize transcript abundance such that, in these examples, predicted risk of IBD would be highest in individuals with high and low expression of *SYNGR1* and *DAP*, respectively. Summing *z* scores over all contributing transcripts identified as the targets of eQTLs in blood, the TRS was correlated with the GRS but suggested that different individuals had the highest risk of disease (**Fig. 1c**).

A TRS based on all 157 candidate genes ascertained from the Blood eQTL browser distinguished individuals with Crohn's disease from control individuals (Δs.d. = 0.51, $P = 0.0019$; **Supplementary Fig. 2a**), but with just a slight improvement on the performance of a classical weighted-allelic-sum GRS based on the very same IBD-associated SNPs that also have eQTL activity (Δs.d. = 0.51, $P = 0.02$). However, this set might have included some genes at which the eQTL action by the GWAS SNP does not necessarily imply pathogenicity (being instead due, for example, to pleiotropy or linkage). Several recent methods such as coloc and summary-data-based Mendelian randomization (SMR) have been developed to ask in a formal statistical framework whether independent signals are consistent with the same variant producing the signals in both studies[15–17]. We ran coloc[15] for all 157 associated candidate genes (Online Methods) and prioritized 29 genes that had the strongest evidence for colocalization of association signals ($H_4 > 80\%$ using GWAS *P* values for Crohn's disease, ulcerative colitis and IBD; **Supplementary Fig. 3** and **Supplementary Table 3**).

The high-confidence set of 29 candidate genes excelled at distinguishing disease status (**Fig. 2a**) as well as progression to complicated disease, namely stricturing (B2) or penetrating/fistulizing (B3) disease according to the Montreal classification system (**Fig. 2b**). The TRS distribution of Crohn's disease samples was highly significantly greater than that of individuals without IBD, who fell almost entirely below the mean risk score of the cases (Δs.d. = 1.46, $P = 1 \times 10^{-13}$). Similarly, the small group that progressed to complicated disease showed significantly higher scores than individuals who remained in the milder B1 state (Δs.d. = 0.63, $P = 5 \times 10^{-5}$). Notably, this discrimination appeared regardless of tissue inflammation, as inflamed and non-inflamed B1 samples had similar TRSs (**Supplementary Fig. 4**). To ensure the robustness of these observations, we repeated the analyses on the basis of a partially overlapping set of 39 genes detected by SMR as targets of IBD-associated variants (Bonferroni-adjusted $P_{SMR} < 2.3 \times 10^{-4}$, 5% Bonferroni; Online Methods and **Supplementary Table 4**). This larger list of genes rendered similar results, distinguishing again between B1 and B2/B3 disease behavior (TRS: Δs.d. = 0.44, $P = 0.007$; **Supplementary Fig. 5a,b**), confirming the power of TRSs.

In contrast, none of the comparisons rendered significant differences when using the corresponding GRSs based on GWAS-associated SNPs (for example, using the loci ascertained by coloc; **Fig. 2a,b**). Furthermore, genome-wide polygenic risk scores (PRSs) assessed using LD pruning[8] across the full range of inclusion thresholds failed to approach the performance of the TRS, peaking at Δs.d. = 0.69 and $P = 9 \times 10^{-4}$ for 668 SNPs at $P < 0.001$ for the disease–control comparison (**Supplementary Fig. 6**). Consistent with recent GWAS results indicating independent genetic contributions to susceptibility and prognosis in Crohn's disease[18], no PRS approached significance for disease progression, which further highlights the enhanced resolution provided by TRSs.

The above results are based on ileal gene expression profiles but use eQTLs that are likely enriched for immune functions, as they were detected in blood from healthy adults. Applying the approach to an independent sample of gene expression in peripheral blood, the TRS also distinguished 61 pediatric Crohn's disease cases and 12 controls (Δs.d. = 1.2, $P = 4 \times 10^{-5}$). We next hypothesized that ileal mucosal samples might include effects that are not observed in peripheral blood but can be important for IBD pathology and are thus likely to improve the power of TRSs. eQTL mapping in 365 RISK samples identified associations at $P < 1 \times 10^{-5}$ for 40 SNPs with 46

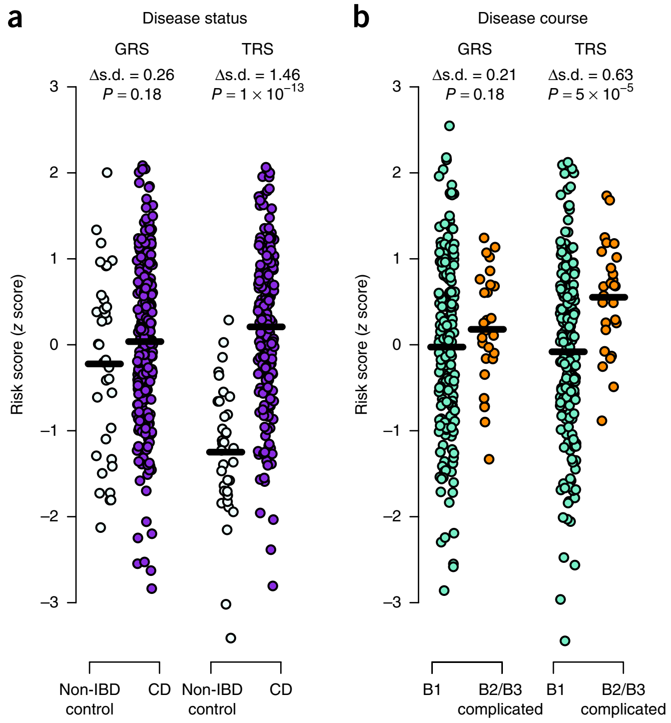

Figure 2 Transcriptional risk scores based on ileal gene expression at diagnosis distinguish status and course of Crohn's disease. (**a**,**b**) A total of 29 genes were predicted by coloc to be the targets of the associations with IBD discovered by GWAS. In contrast to classical GRSs based on allele counts, risk scores based on summation of standardized expression of these IBD-associated genes (TRSs), after polarization according to direction of risk, distinguish between individuals with Crohn's disease (CD; $n = 210$) and controls ($n = 35$) (**a**) and between individuals with Crohn's disease who remain in B1 ($n = 183$) and those who go on to develop complicated disease (B2 and/or B3; $n = 27$) (**b**) within 3 years of diagnosis. Differences between groups (in s.d. units) along with P values (two-sided t test) are reported for each comparison.

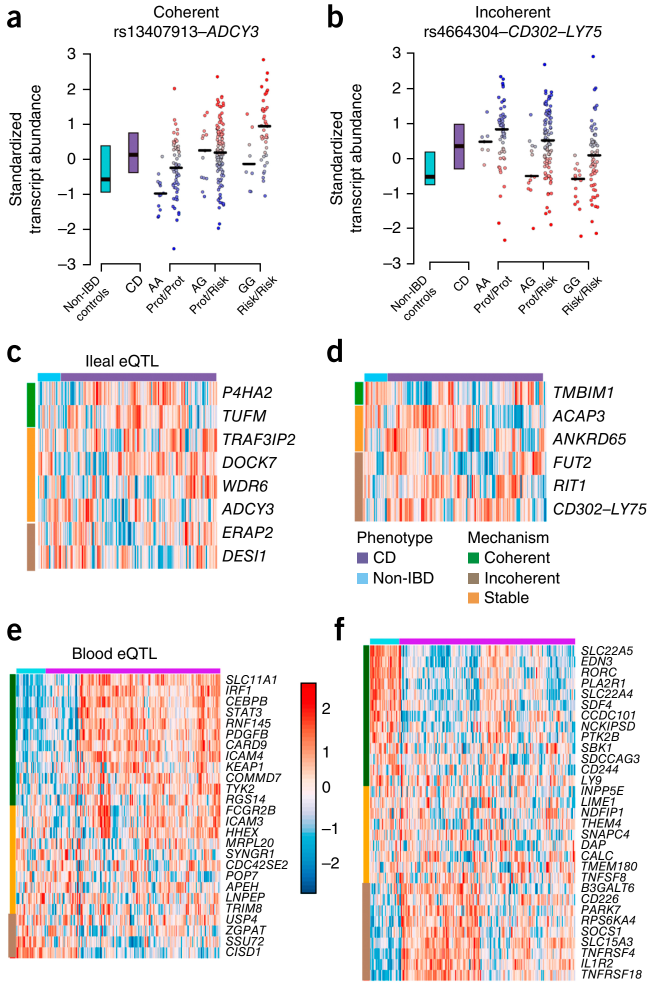

Figure 3 Gene expression polarized according to predicted direction of risk uncovers two divergent mechanisms of association with disease. For about half of the eQTLs, *trans* and environmental effects result in coordinated modification of gene expression in cases relative to controls. (**a**) Example of a coherent association, where individuals with the risk genotype (GG) show increased expression of *ADCY3*, consistent with the prediction based on the direction of effect of this allele as an eQTL in ileal tissue. Left and right columns of individual points for each genotype correspond to cases ($n = 210$) and controls ($n = 235$), respectively. The purple and light blue boxes depict the median and interquartile range for each group. (**b**) Example of an incoherent association, where individuals with the risk allele have reduced expression in the opposite direction to the overall increased levels of *CD302–LY75* in cases. (**c**,**d**) Considering eQTLs discovered in ileal tissue, eight genes are controlled by ileal eQTLs that increase their expression (**c**) and six genes are controlled by eQTLs that decrease their expression (**d**). Purple and light blue bars above the heat maps indicate cases ($n = 210$) and controls ($n = 35$), respectively; bars along the left indicate genes that are coherent (green), incoherent (red) and stable (orange) with respect to disease. (**e**,**f**) Considering eQTLs discovered in blood, 26 genes are upregulated (**e**) and 31 genes are downregulated (**f**) by the allele associated with IBD. In this case, 25 genes are coherent and just 13 are incoherent. The heat map is color-indexed according to the z score of each gene from low (blue) to high (red) expression.

genes that fell in the vicinity (within 1 Mb) of the 232 SNPs associated with IBD (Online Methods and **Supplementary Table 5**). These included associations known to be active in ileal tissue such as *FUT2* and *ERAP2* (refs. 14,19,20).The list of ileum effects included 27 genes not described in the Blood eQTL browser, 7 of which were selected by coloc as having joint eQTL and GWAS effects consistent with a causal contribution to IBD ($H_4 > 80\%$ for the three phenotypes considered; **Supplementary Table 5**). A TRS based on this short list of seven loci, using the direction of effect of each eQTL in ileum to polarize risk, failed to separate samples according to disease status (Δs.d. = 0.17, $P = 0.32$) or course of disease (Δs.d. = −0.11, $P = 0.53$). Surprisingly, a 14-gene TRS including 7 more ileum-specific loci exclusively detected by SMR also failed to discriminate cases and controls.

In addition to *cis* effects, gene expression is also influenced by a combination of *trans*-acting genetic effects and environmental effects, both of which tend to produce coordinated patterns of gene expression that may disrupt the expected coherence of the signs of eQTL and GWAS effects[21,22]. Specifically, IBD pathology is accompanied by altered expression of many genes as a response to altered intestinal microbiota[23,24]. For example, **Figure 3a** shows how *ADCY3* is upregulated in individuals with Crohn's disease, consistent with the direction of the eQTL effect shown by IBD risk allele rs13407913[G] ($\beta = 0.14$, $P = 4 \times 10^{-16}$), whereas *CD302–LY75* is induced in the mucosa of patients with Crohn's disease despite being downregulated by the GWAS risk allele rs4664304[G] ($\beta = -0.065$, $P = 4 \times 10^{-7}$; **Fig. 3b**). Detailed exploration on a gene-by-gene basis (**Fig. 3c**,**d**) suggests that this type of disruption may account for the poor performance of the 14-gene TRS based on ileum eQTL effects. The three genes acting in a coherent fashion (Δs.d. > 0.3 between cases and controls in the

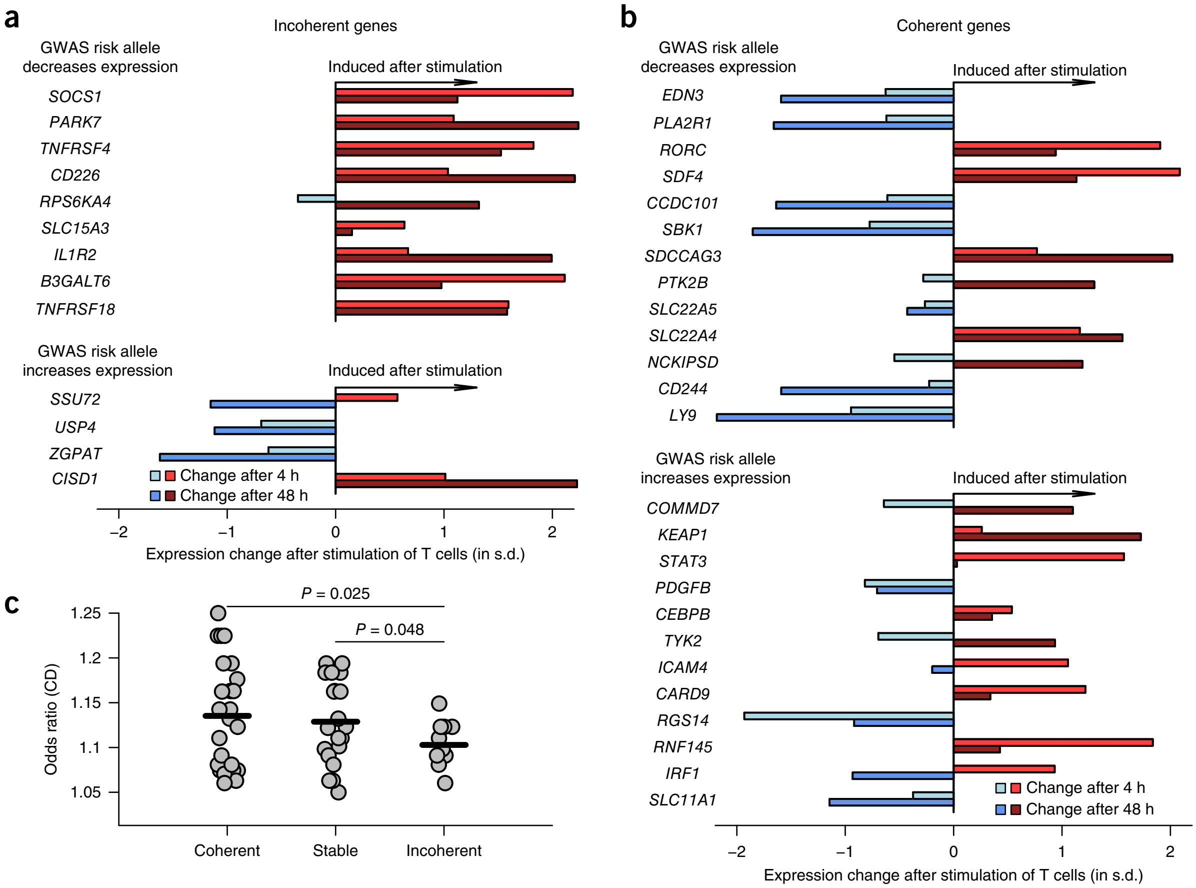

Figure 4 Incoherent genes show similar patterns in stimulated immune cells and are more weakly associated with IBD according to GWAS. The data set includes changes in gene expression after 4 h and 48 h in primary T cells stimulated with anti-CD3/CD28 beads as reported by the ImmVar consortium. (**a**) All but one of the 13 incoherent genes show changes in expression at 48 h that mimic the inconsistent tendencies observed in individuals with Crohn's disease from the RISK cohort. (**b**) Coherent genes show more diverse changes in patterns of expression. (**c**) Incoherent genes ($n = 13$) have significantly lower odds ratios of association with IBD by GWAS than coherent ($n = 25$) or stable ($n = 19$) genes. P values (two-sided t test) are reported for each pairwise comparison.

predicted direction) enhance TRS performance, but they are offset by the five genes whose incoherence (Δs.d. > 0.3 in the opposite direction) diminishes the performance of the TRS. The other six genes are stable with respect to disease status, not showing a significant difference in expression between cases and controls.

By contrast, for the 57 genes detected by either coloc or SMR as target genes on the basis of eQTL effects in blood, there was a clear excess of coherent associations ($n = 25$) over incoherent ones ($n = 13$) (**Fig. 3e**,**f**). Clearly, most of the coherent and incoherent genes are strongly co-regulated, implying that environmental or other *trans* effects mediate the paradoxical deviation between observed and predicted directions of effect, rather than confounding effects of secondary *cis*-acting alleles. Examples of incoherence include *CD226*, encoding an immunoglobulin receptor involved in control of viral infection[25] and implicated in several autoimmune diseases[26], which is induced in individuals with Crohn's disease (Δs.d. = 1.07) in spite of being downregulated by the GWAS risk allele rs727088[G] ($P = 1 \times 10^{-46}$; **Supplementary Table 3**). Similarly, *TNFRSF18* encodes a receptor of the tumor necrosis factor (TNF) family with a key role in maintaining self-tolerance[27,28] and is also induced in individuals with Crohn's disease (Δs.d. = 1.49) even though the risk allele decreases its expression (**Supplementary Table 3**). The functional evidence for both genes suggests a scenario in which induction is protective (for example, to clear infection in the gastrointestinal tract); hence, individuals with the GWAS risk allele are more prone to developing chronic inflammation because they fail to induce expression sufficiently to fully engage the defense response.

Consistent with this interpretation, analysis of ImmVar consortium data on *ex vivo* responses to 4 h and 48 h of stimulation[29] indicated a common theme for the 13 incoherent genes. The nine genes that were incoherently upregulated in individuals with Crohn's disease were also induced in CD4$^+$ T cells after 48 h of stimulation with anti-CD3/CD28 beads, whereas three of the four genes that were incoherently downregulated in affected individuals were also suppressed after immune stimulation (**Fig. 4a**). The coherently regulated genes did not show such a consistent pattern (**Fig. 4b**), suggesting that their disease response may not be due to immune stimulation. This difference between the two sets of genes was significant ($P = 0.03$, Fisher's exact test), and similar results applied to the effects of stimulation with lipopolysaccharide (LPS) or infection with influenza virus (data not shown).

Overall, the contrasting behavior of coherent and incoherent genes is consistent with the notion that gene-regulatory IBD risk alleles have detrimental effects through two different mechanisms: some directly promote disease because they regulate gene expression in a manner that is inherently pathogenic, and others fail to safeguard individuals by insufficiently engaging protective shifts of gene expression.

Intriguingly, the latter class generally has odds ratios around 1.1, which is significantly lower than for the remainder (**Fig. 4c**). Biopsy gene expression profiling of larger cohorts should confirm this inference and further refine our ability to distinguish active and protective risk mechanisms. Other interpretations are also possible, including the possibility that eQTL effects in the ileum are not contributing strongly to pathogenesis and processes unique to individual genes. An excellent example of the latter is the one incoherent gene that contravenes our model, *CISD1*, which encodes mitoNEET, an Fe/S-domain protein localized to the mitochondria where it is required for redox sensing[30]. Mitochondrial function is protective against progression in Crohn's disease[4,31], yet transcription of *CISD1* was downregulated in patients overall and strongly induced in T cells by *ex vivo* stimulation, and the risk allele increased expression.

The existence of incoherent associations highlights the fact that there is much to learn about the relationship between eQTL effects and disease pathogenesis. This phenomenon is likely also to apply to other autoimmune and inflammatory diseases, and further dissection should in turn improve the development of TRSs that are predictive of progression to complicated disease, with implications for therapeutic treatment.

URLs. Blood eQTL browser, http://genenetwork.nl/bloodeqtlbrowser/; GTEx, http://www.gtexportal.org/home/; IIBDGC trans-ancestry meta-analysis association data, https://www.ibdgenetics.org/downloads.html; SMR, http://cnsgenomics.com/software/smr/index.html; coloc R package, https://cran.r-project.org/web/packages/coloc/index.html; 1000 Genomes Project, http://www.internationalgenome.org/1000-genomes-browsers; GEO, https://www.ncbi.nlm.nih.gov/geo/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to B. Zeng, D. Arafat, H. Somineni, S. Venkateswaran, and colleagues from the Gibson and Kugathasan laboratories for their support and helpful comments. We also would like to thank I. Mendizabal, J. Lachance and K. Jordan for comments on the manuscript. This research was supported by Project 3 (G.G., PI) of the NIH program project "Statistical and Quantitative Genetics" grant P01-GM0996568 (B. Weir, University of Washington, Director) as well as research grants from the Crohn's and Colitis Foundation of America (CCFA), New York, to the individual study institutions participating in the RISK study.

AUTHOR CONTRIBUTIONS

U.M.M. and G.G. conceived the theoretical framework for the TRSs. L.A.D., J.S.H. and S.K. participated in the conception and design of the RISK study. K.M., J.P., T.D.W., A.G., J.D.N., W.V.C., J.R.R., D.R.M., R.K., M.B.H., S.S.B., M.C.S., R.N.B., J.F.M., M.C.D., B.J.A., M.-O. K. and J.C. recruited subjects, collected the data, and worked on its curation and analysis. U.M.M. performed the TRS analyses. U.M.M. and G.G. interpreted the results and drafted the manuscript, while L.A.D., J.S.H. and S.K. assisted with results interpretation and writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Fairfax, B.P. & Knight, J.C. Genetics of gene expression in immunity to infection. *Curr. Opin. Immunol.* **30**, 63–71 (2014).
2. Gibson, G., Powell, J.E. & Marigorta, U.M. Expression quantitative trait locus analysis for translational medicine. *Genome Med.* **7**, 60 (2015).
3. Haberman, Y. *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Invest.* **124**, 3617–3633 (2014).
4. Kugathasan, S. *et al.* Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *Lancet* **389**, 1710–1718 (2017).
5. Witte, J.S., Visscher, P.M. & Wray, N.R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15**, 765–776 (2014).
6. Wray, N.R., Yang, J., Goddard, M.E. & Visscher, P.M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
7. Wray, N.R., Goddard, M.E. & Visscher, P.M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
8. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
9. Walters, T.D. *et al.* Increased effectiveness of early therapy with anti–tumor necrosis factor-α vs an immunomodulator in children with Crohn's disease. *Gastroenterology* **146**, 383–391 (2014).
10. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
11. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
12. Kabakchiev, B. & Silverberg, M.S. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology* **144**, 1488–1496 (2013).
13. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
14. Di Narzo, A.F. *et al.* Blood and intestine eQTLs from an anti-TNF-resistant Crohn's disease cohort inform IBD genetic association loci. *Clin. Transl. Gastroenterol.* **7**, e177 (2016).
15. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
16. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
17. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
18. Lee, J.C. *et al.* Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat. Genet.* **49**, 262–268 (2017).
19. Ning, K. *et al.* Improved integrative framework combining association data with gene expression features to prioritize Crohn's disease genes. *Hum. Mol. Genet.* **24**, 4147–4157 (2015).
20. Singh, T. *et al.* Characterization of expression quantitative trait loci in the human colon. *Inflamm. Bowel Dis.* **21**, 251–256 (2015).
21. Albert, F.W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
22. Gibson, G. & Weir, B. The quantitative genetics of transcription. *Trends Genet.* **21**, 616–623 (2005).
23. de Souza, H.S. & Fiocchi, C. Immunopathogenesis of IBD: current state of the art. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 13–27 (2016).
24. McGovern, D.P., Kugathasan, S. & Cho, J.H. Genetics of inflammatory bowel diseases. *Gastroenterology* **149**, 1163–1176 (2015).
25. Nabekura, T. *et al.* Costimulatory molecule DNAM-1 is essential for optimal differentiation of memory natural killer cells during mouse cytomegalovirus infection. *Immunity* **40**, 225–234 (2014).
26. Martinet, L. & Smyth, M.J. Balancing natural killer cell activation through paired receptors. *Nat. Rev. Immunol.* **15**, 243–254 (2015).
27. Petrillo, M.G. *et al.* GITR$^+$ regulatory T cells in the treatment of autoimmune diseases. *Autoimmun. Rev.* **14**, 117–126 (2015).
28. Reikvam, D.H. *et al.* Increase of regulatory T cells in ileal mucosa of untreated pediatric Crohn's disease patients. *Scand. J. Gastroenterol.* **46**, 550–560 (2011).
29. Ye, C.J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
30. Wiley, S.E. *et al.* The outer mitochondrial membrane protein mitoNEET contains a novel redox-active 2Fe-2S cluster. *J. Biol. Chem.* **282**, 23745–23749 (2007).
31. Novak, E.A. & Mollen, K.P. Mitochondrial dysfunction in inflammatory bowel disease. *Front. Cell Dev. Biol.* **3**, 62 (2015).

ONLINE METHODS

Cohort and outcome classification. The RISK study is an observational prospective cohort study that aims to develop risk models for predicting complicated course in children with Crohn's disease. From 2008 to 2012, the RISK study recruited more than 1,800 treatment-naive patients with a suspected diagnosis of Crohn's disease at 28 pediatric gastroenterology centers in North America[3,4]. This disorder is a chronic inflammatory condition of the gastrointestinal tract that results from inappropriate activation of the immune system thought to be due to a combination of host genetic makeup, enteric flora, and microbial or other pathological triggers. A minority of patients progress with time to complicated disease that may require surgery and/or intensive pharmacological therapy. We used the Montreal criteria to classify patients according to disease behavior, distinguishing non-complicated B1 disease (non-stricturing, non-penetrating disease) from complicated disease, composed of B2 (stricturing) and/or B3 (penetrating) behavior[32,33].

We ascertained 245 samples from the RISK study that had been profiled with ileal RNA-seq and genotyped with the Illumina high-density Immunochip array. 35 of the ascertained individuals lacked gut inflammation and were classified as non-IBD controls. The remaining selected individuals showed persisting Crohn's disease and remained in complication-free B1 status for at least 90 d from the time of initial diagnosis. After 3 years of follow-up, 17 and 10 patients progressed to B2 and B3 status, respectively. We joined the latter 27 samples to form a group of patients with complicated disease course. The majority of individuals were of European ancestry ($n = 210$; 85.7%), with smaller fractions of samples with African ($n = 10$; 4.1%) and other/mixed ($n = 25$; 10.2%) ancestry. More details about outcome classification are available in Kugathasan *et al.*[4].

Along with disease behavior, disease location has a key role in the natural history and clinical course of patients diagnosed with Crohn's disease. Because a recent study showed that a GRS for IBD could distinguish patients with ileal/ileocolonic disease from those with only colonic disease[34], we asked whether a TRS could also distinguish these two classes of Crohn's disease. According to the Paris modification[32] of the Montreal classification, pediatric disease is also classified into L1 (ileal only), L2 (colonic only), L3 (ileocolonic) and L4 (upper gastrointestinal tract). For **Supplementary Figure 4**, we combined L1 and L3 into inflamed B1, as the biopsies were taken from the ileum, whereas L2 was uninflamed relative to the site of biopsy. No L4 cases were available. The analysis confirmed that the TRS indeed distinguished L1/L3 from L2 disease.

Processing of RNA-seq data from ileal biopsies and SNP data from the RISK cohort. RNA was isolated from ileal biopsies obtained from colonoscopy at diagnosis, and profiles of gene expression were determined using RNA-seq as previously reported. Reads were mapped to the human genome (hg19) with TopHat 2.0.13 using default parameters[35]. Aligned reads were transformed with SAMtools[36] to quantify the number of reads at the gene level with HTSeq-0.6.1 (ref. 37) using default "union" mode. Raw counts were compiled and processed with edgeR[38] to obtain normalized counts through trimmed mean of *M*-values normalization. An in-house R script was then used to inverse rank transform expression estimates for each gene into a standard normal distribution with mean 0 and variance 1. For comparison with GTEx, the data were further transformed into the reads per kilobase per million mapped reads (RPKM) metric[39], and 13,769 genes with RPKM >1 and >6 read counts in at least ten individuals were retained. The median RPKM per gene in RISK and the median RPKM per gene in 53 tissues available from GTEx (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_median_rpkm.gct; see URLs) had a median Spearman correlation of 0.57 (range 0.39–0.88), with the largest correlations corresponding to GTEx "Small_Intestine.Terminal_Ileum" ($r_s = 0.88$), "Colon.Transverse" ($r_s = 0.79$) and "Stomach" ($r_s = 0.72$), confirming similarity of the RISK biopsy data to an external bowel data set.

The Immunochip was designed to densely genotype 186 distinct loci containing markers associated at genome-wide significance ($P < 5 \times 10^{-8}$) with 12 autoimmune and inflammatory diseases, including Crohn's disease and ulcerative colitis. The array was designed to contain all 1000 Genomes pilot phase (September 2009 release) SNPs within 0.1-cM recombination blocks (HapMap 3 CEU) around the top associated markers by GWAS[40]. Initial calling of the Immunochip array before quality control contained 192,523 variants. We used the Bioconductor SNPlocs.Hsapiens.dbSNP.20120608 package[41] to map autosomal SNPs to GRCh37 and remove (i) non-biallelic variants, (ii) SNPs not in Hardy–Weinberg equilibrium ($P < 1 \times 10^{-3}$) and (iii) variants not present in the 1000 Genomes Phase 1 variant set (March 2012 release). At this point, there were 161,540 remaining SNPs. We further removed 49,253 variants with MAF <5% and 10,874 SNPs with missing data rate >1% across all individuals. After quality control, there were 101,413 genotyped variants available for analysis, and all 245 individuals presented genotype missing rates <0.1%. To check relatedness among samples, we calculated pairwise identity by descent based on 26,233 SNPs obtained after LD pruning using the PLINK routine "--indep 50 5 2," confirming minimal overall relatedness (PI_HAT < 0.05 for 99.3% of pairwise comparisons) with just three pairs of first-degree relatives (PI_HAT > 0.25).

Selection of SNPs and candidate genes associated with IBD by GWAS. Because our goal was to uncover genes involved in susceptibility to Crohn's disease, we considered as candidates all genes with a transcription start site (TSS) located ±1 Mb with respect to each of the 232 independent GWAS SNPs previously associated with IBD[10]. We examined 7,389 SNP–gene pairs, including 6,180 unique candidate genes (32 genes considered per SNP on average, range of 5 to 620 genes). The Blood eQTL browser (see URLs) was queried to ascertain which genes are under the control of IBD-associated SNPs. We observed 163 instances in which the GWAS SNP ($n = 129$) or a SNP in LD ($n = 34$, at $r^2 > 0.8$ in 1000 Genomes CEU data) acted as an eQTL (FDR < 5%) for a candidate gene located <1 Mb from the associated SNP (**Supplementary Table 1**). In total, this resulted in selection of 157 unique genes (6 genes were under the control of two different IBD SNPs).

Mapping study in RISK cohort to build the ileal TRS. A fraction of eQTL variants are known to act in a tissue-specific manner[13]. We used the RISK ileal biopsies to perform a targeted eQTL study focused on the 7,389 SNP–gene pairs. This analysis aimed to confirm whether eQTLs discovered in peripheral blood are also present in ileal tissue and to detect ileal-specific eQTLs that can be used to pinpoint new candidate pathogenic genes.

We applied several quality control steps to remove batch effects and normalize the matrix of gene expression to carry out the eQTL mapping study. First, we performed a sex incompatibility check comparing the sex recorded for each individual to the expression of *XIST* and Y-chromosome genes *EIF1AY*, *RPS4Y1*, *DDX3Y* and *KDM5D*. A heat map based on expression of these five genes did not show any sex mismatch. Next, we tried to identify low-quality samples using *D* statistics as done by GTEx[13]. For each sample, mean correlation of expression with the remaining samples was calculated. All samples showed D >0.9 with no obvious visual outliers from the average correlation of 0.972, and all samples were therefore kept for further analysis.

Finally, supervised normalization procedures were used to remove global effects present in the matrix of expression data. The transcriptome shows pervasive co-regulation of transcript abundance that leads to modules of co-regulated genes that have similar biological functions[42]. Biological variables such as disease can also induce massive changes in gene expression (for example, thousands of genes are differentially expressed among groups in the RISK study[4]). Moreover, hidden batch effects and other unknown cofounders can induce spurious correlations at the genome-wide level. All these sources of biological and/or technical variability can hamper the detection of locally acting *cis*-eQTLs. We first used unsupervised surrogate variable analysis (SVA)[41] to identify hidden confounding factors, deliberately protecting known variables such as sex and disease status (to be included as covariates in the eQTL mapping step). The algorithm detected 14 surrogate variables that were removed using the supervised normalization of microarray (SNM) procedure[42]. Specifically, we fit sex and disease status as biological variables and removed the effects of the 14 estimated surrogate variables by including these as adjustment variables with the rm = T option.

eQTL mapping was performed using a linear mixed model implemented in GEMMA[43], which allows adjustment for population structure and relatedness among individuals as a random effect through a genetic relationship matrix (GRM) based on the LD-pruned SNP data set. We tested for associations between genotype and normalized gene expression, including sex and disease status as covariates. **Supplementary Table 2** reports association results for 136 available SNP–gene pairs (ileal eQTL association data were not available for the remaining 21 pairs).

doi:10.1038/ng.3936

Gene selection with SMR and coloc. Detection of nominally significant associations both for eQTLs and with IBD at a single SNP does not necessarily imply that the SNP is responsible for both effects. Several recent methods have been designed to increase confidence that colocalization of signals implies that the gene affected by the regulatory SNP is also responsible for the trait association. coloc uses a Bayesian framework to infer whether the two signals are due to a single site or to two sites in LD within a genomic region of interest[15]. It calculates posterior probabilities to quantify the support for five different hypotheses regarding the presence and sharing of causal variants by the two traits under consideration. Similarly, SMR combines GWAS and eQTL summary association data to prioritize target genes with evidence for causal or pleiotropic effects[17]. We applied both methods to ascertain target genes from the list of 157 aforementioned candidate genes.

We used GWAS summary data for Crohn's disease, ulcerative colitis and IBD from the publicly available IIBDGC GWAS plus Immunochip trans-ancestry MANTRA meta-analyses (see URLs). For each of the three disease phenotypes, we processed the data considering the sample size indicated in Table 1 of Liu *et al.*[10]. For eQTL effects, we used the *cis*-eQTL summary data from the largest existing immune-related data set, namely the Blood eQTL browser (see URLs), and converted the reported z statistics into β and standard error values following the guidelines from the SMR Supplementary Note in Zhu *et al.*[17]. The assigned sample size was 5,311, using Europeans from the 1000 Genomes Project as the reference sample for MAF and LD patterns (see URLs). For the coloc analyses, we considered as validated target genes 29 independent loci with 80% or greater posterior probability of the hypothesis of one causal variant common to both traits (H_4) for all three of the phenotypes. For the SMR analyses, **Supplementary Figure 3a** shows the strong relationship between the SMR P value and highly significant P values for both the GWAS and eQTL effects. This validates the selection of loci (such as the red and brown dots in the figure) that passed Bonferroni correction for all three of the phenotypes considered (significance threshold $P < 2.3 \times 10^{-4}$ for one phenotype, as the P values are highly correlated). However, it also highlights the likely dependence of the SMR statistic on the significance of the eQTL effects, which in turn are strongly influenced by the sample size, as noted by Zhu *et al.*[17]. In general, inclusion of more high-confidence genes would be expected to improve the TRS in part by reducing the variance of the score, and it is therefore likely that the small sample size for the ileal eQTL results contributes to its weaker diagnostic performance relative to the larger blood-derived gene set. We also replicated the case–control comparison with an analysis of 13 of the 26 genes recently reported from immune cell-type-specific eQTLs[44] for which replicated directional effects could be inferred (Δs.d. = 0.73, $P = 3 \times 10^{-5}$), but, again, larger sample sizes will be needed to establish a high-confidence set of such genes. Owing to the low density of variants on the Immunochip and the likely presence of multiple causal effects at each locus, computation and interpretation of SMR's HEIDI scores was compromised for half of the loci; as only four were inferred to be unambiguously causal by this test ($P < 2.3 \times 10^{-4}$ for the three phenotypes), it was deemed not useful for selection of genes for TRS computation. 15 of the loci are common to SMR and coloc, implying that the methods are complementary. Summary results for all genomic regions considered are available in **Supplementary Tables 3** and **4**.

In addition, we used coloc to select causal genes among the 27 genes controlled by ileum-specific eQTLs ($P < 1 \times 10^{-5}$) discovered in the mapping study described above. To do so, we extended the eQTL mapping study ±500 kb around the susceptibility SNP for each of the genomic regions and processed the association data to run coloc and SMR on these regions. We considered as validated seven target genes that showed H_4 >80% for all three of the disease phenotypes. Because of the low number of loci detected through coloc, we complemented the analyses with seven more loci that passed SMR for all three phenotypes (Bonferroni-adjusted $P < 0.00185$ inclusion threshold for one phenotype). Summary results are available in **Supplementary Table 5**.

Calculation of GRSs and TRSs. We carried out several comparisons to contrast the predictive power of the TRS with that of the GRS based on the corresponding GWAS SNPs (those that act as eQTLs for the selected genes). For the GRS, we used the "score" routine available in PLINK to generate a GRSs weighted using the log(OR) for IBD from GWAS meta-analysis[10] (reported in **Supplementary Table 1**; weighting by the log(OR) for Crohn's disease rendered very similar scores at each comparison). In turn, calculation of the TRS consisted of three steps. First, we used the eQTL activity of GWAS SNPs to infer the direction of risk at each gene selected for the TRS. We used "High Expr." and "Low Expr." (available in **Supplementary Tables 1–5**) to denote whether the risk allele associated with disease led to increased (High Expr.) or decreased (Low Expr.) gene expression. Next, we polarized expression values so that elevated risk, irrespective of the sign of the effect on expression, added to the TRS. This was done simply by changing the sign of the z score for genes labeled as Low Expr. (for example, expression z scores of −1.5 and +1.3 would transform into +1.5 and −1.3, respectively). Finally, we obtained the TRS for each individual by summing the polarized z scores over all genes and rank normalizing the distribution. We used t tests to compare the performance of the GRS and TRS between groups.

Calculation of PRSs. PRSs have emerged as the gold standard for overall prediction from GWAS. We used the P+T (pruning + threshold)[8] method to build PRSs based on independent SNPs that passed different significance thresholds in GWAS analysis. To avoid loss of power due to the inclusion of correlated SNPs, we first selected 15,135 LD-pruned SNPs from the RISK Immunochip data (by running PLINK's indep-pairwise routine on 5,000 randomly selected individuals from the UK Biobank). Then, we used PLINK's score routine to calculate a battery of PRSs based on variants selected across the complete spectrum of significance thresholds for inclusion (from 329 SNPs at $P < 0.00001$ to 9,214 SNPs at $P < 0.5$) in the IIBDGC GWAS plus Immunochip trans-ancestry MANTRA meta-analyses for IBD (see URLs). The performance of the PRSs for both the case–control comparison and the indolent disease–complicated disease comparison at different thresholds is reported in **Supplementary Figure 6**. The performance of the PRSs between groups was tested through t tests.

Coherence and incoherence. For the evaluation of coherence between eQTL and disease effects, we first evaluated whether each transcript was significantly differentially expressed between control and Crohn's disease samples by at least 0.3 s.d. units (P ~0.05). Despite the small sample size of controls, clear co-regulation of the upregulated (**Fig. 3c**,**e**) and downregulated (**Fig. 3d**,**f**) genes is clearly visualized. Next, we classified as coherent genes for which the direction of the eQTL effect was the same as the effect for the disease (that is, increased expression of the risk allele as well as elevated expression in cases relative to controls or decreased expression of the risk allele and repressed expression in cases). Incoherent genes were those with the opposite relationship (that is, either increased expression of the risk allele and repression in cases or vice versa). Stable genes were those without clear differences in expression between cases and controls.

Whereas our initial proposal for the TRS assumed no global impact of disease on gene expression[2], the RISK data set showed that fewer than half of the candidate genes were stable by the above definition. Coherence mathematically tended to enhance the performance of the TRS as it elevated the difference between cases and controls for each gene. By contrast, incoherence diminished TRS performance as the polarized eQTL effect was counteracted by the influence of disease. Because there was an excess of incoherent associations for ileal eQTLs, the TRS performance was compromised. However, as there was no global difference in expression of the GWAS candidate genes between B1 cases and complicated B2/B3 cases, coherence and incoherence did not affect the ability of the TRS to discriminate these conditions.

Functional evidence from the ImmVar project. We used data from the ImmVar project (GEO accession GSE60235) to gain further insight into the coherent and incoherent behaviors detected for some genes included in the TRS. The data set includes expression profiling with the Affymetrix Human Gene 1.0 ST array of resting and activated T cells from 15 healthy human individuals collected under five different conditions[29]. We downloaded the matrix of normalized gene expression and selected experiments corresponding to three conditions, namely "Unstimulated 4hr" (n = 15), "Activated 4hr" (n = 15) and "Activated 48hr" (n = 15). For each gene of interest, we transformed expression estimates into a standard normal distribution with mean 0 and variance 1 and performed pairwise comparisons to explore the changes in gene expression at 4 h and 48 h after stimulation with anti-CD3 and anti-CD28 beads. The changes in average z score for the selected genes are reported in **Figure 4**. We observed similar patterns for both coherent and incoherent genes analyzing a similar ImmVar project that profiled changes in monocyte-derived dendritic cell gene expression after stimulation with LPS or influenza (GEO accession GSE53166)[45].

Data availability. The RNA-seq data for the 245 individuals included in this study have been deposited in the Gene Expression Omnibus (GEO) and are accessible through GEO series accession GSE93624. A **Life Sciences Reporting Summary** is available.

32. Levine, A. *et al*. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm. Bowel Dis.* **17**, 1314–1321 (2011).
33. Satsangi, J., Silverberg, M.S., Vermeire, S. & Colombel, J.F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* **55**, 749–753 (2006).
34. Cleynen, I. *et al.* Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* **387**, 156–167 (2016).
35. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
36. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
38. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
39. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
40. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
41. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
42. Mecham, B.H., Nelson, P.S. & Storey, J.D. Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
43. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
44. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
45. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).

nature research

Corresponding author(s): Greg Gibson

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

No sample size calculation was performed. We used the samples available from the RISK study, an inceptional prospective cohort without a predicted target regarding sample size (all pediatric individuals with symptoms of Crohn's disease were sampled at 28 clinics over a period of three years).

2. Data exclusions

Describe any data exclusions.

No data or sample was excluded. We used the 245 samples from the RISK study that had i) RNA-Seq available, ii) SNP data available and iii) had been profiled over 3 years (as described in Kugathasan et al., Lancet, 2017)

3. Replication

Describe whether the experimental findings were reliably reproduced.

We performed a single replication attempt, as described in the main text ("Applying the approach to an independent sample of peripheral blood gene expression, the TRS also distinguished 61 pediatric Crohn's disease cases and 12 controls (ΔSD=1.2; P=4×10-5)."

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples were allocated into three different groups: i) non-IBD controls; ii) CD patients that remain in B1 status over a three year period and iii) CD patients that developed either B2 and/or B3 complications in the window from 90-days after diagnosis to 3-year after diagnosis. No covariates were controlled for. The details are available in the 1st/2nd paragraph of the Online Methods.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not considered, given that group allocation had been previously done previous to this study by the investigators from the RISK study (Kugathasan et al. Lancet, 2017)

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	For all our analyses we used the R software environment for statistical computing

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All used samples are available for public use (SNP data is available upon contact with corresponding author)

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used
b. Describe the method of cell line authentication used.	No eukaryotic cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	No eukaryotic cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No eukaryotic cell lines were used

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	All used samples belong to the cases of pediatric Crohn's disease previously reported in Kugathasan et al. (Lancet, 2017).