Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis

Urmo Võsa*#[1,2], Annique Claringbould*#[1], Harm-Jan Westra**[1], Marc Jan Bonder**[1], Patrick Deelen**[1,3], Biao Zeng[4], Holger Kirsten[5], Ashis Saha[6], Roman Kreuzhuber[7,8], Silva Kasela[2], Natalia Pervjakova[2], Isabel Alvaes[9], Marie-Julie Fave[9], Mawusse Agbessi[9], Mark Christiansen[10], Rick Jansen[11], Ilkka Seppälä[12], Lin Tong[13], Alexander Teumer[14], Katharina Schramm[15,16], Gibran Hemani[17], Joost Verlouw[18], Hanieh Yaghootkar[19], Reyhan Sönmez[20,21], Andrew Brown[22,23,21], Viktorija Kukushkina[2], Anette Kalnapenkis[2], Sina Rüeger[24], Eleonora Porcu[24], Jaanika Kronberg-Guzman[2], Johannes Kettunen[25], Joseph Powell[26], Bernett Lee[27], Futao Zhang[28], Wibowo Arindrarto[29], Frank Beutner[30], BIOS Consortium, Harm Brugge[1], i2QTL Consortium, Julia Dmitreva[31], Mahmoud Elansary[31], Benjamin P. Fairfax[32], Michel Georges[31], Bastiaan T. Heijmans[29], Mika Kähönen[33], Yungil Kim[34,35], Julian C. Knight[32], Peter Kovacs[36], Knut Krohn[37], Shuang Li[1], Markus Loeffler[5], Urko M. Marigorta[4], Hailang Mei[38], Yukihide Momozawa[31,39], Martina Müller-Nurasyid[15,16,40], Matthias Nauck[41], Michel Nivard[42], Brenda Penninx[11], Jonathan Pritchard[43], Olli Raitakari[44], Olaf Rotzchke[27], Eline P. Slagboom[29], Coen D.A. Stehouwer[45], Michael Stumvoll[46], Patrick Sullivan[47], Peter A.C. 't Hoen[48], Joachim Thiery[49], Anke Tönjes[46], Jenny van Dongen[11], Maarten van Iterson[29], Jan Veldink[50], Uwe Völker[51], Cisca Wijmenga[1], Morris Swertz[3], Anand Andiappan[27], Grant W. Montgomery[52], Samuli Ripatti[53], Markus Perola[54], Zoltan Kutalik[24], Emmanouil Dermitzakis[22,23,21], Sven Bergmann[20,21], Timothy Frayling[19], Joyce van Meurs[18], Holger Prokisch[55,56], Habibul Ahsan[13], Brandon Pierce[13], Terho Lehtimäki[12], Dorret Boomsma[11], Bruce M. Psaty[10,57], Sina A. Gharib[58,10], Philip Awadalla[9], Lili Milani[2], Willem Ouwehand[7,59], Kate Downes[7], Oliver Stegle[8,60,61], Alexis Battle[62], Jian Yang[28,63], Peter M. Visscher[28], Markus Scholz[5], Gregory Gibson[4], Tõnu Esko[2], Lude Franke#[1]

* These authors contributed equally to this work.

** These authors contributed equally to this work.

1. Department of Genetics, University Medical Centre Groningen, Groningen, The Netherlands
2. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu 51010, Estonia
3. Genomics Coordination Center, University Medical Centre Groningen, Groningen, The Netherlands
4. School of Biological Sciences, Georgia Tech, Atlanta, United States of America
5. Institut für Medizinische InformatiK, Statistik und Epidemiologie, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
6. Department of Computer Science, Johns Hopkins University, Baltimore, United States of America
7. Department of Haematology, University of Cambridge and NHS Blood and Transplant Cambridge Biomedical Campus, Cambridge, United Kingdom
8. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
9. Computational Biology, Ontario Institute for Cancer Research, Toronto, Canada
10. Cardiovascular Health Research Unit, University of Washington, Seattle, United States of America
11. Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
12. Department of Clinical Chemistry, Fimlab Laboratories and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
13. Department of Public Health Sciences, University of Chicago, Chicago, United States of America
14. Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
15. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
16. Department of Medicine I, University Hospital Munich, Ludwig Maximilian's University, München, Germany

17. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
18. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands
19. Exeter Medical School, University of Exeter, Exeter, United Kingdom
20. Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
21. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
22. Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
23. Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland
24. Lausanne University Hospital, Lausanne, Switzerland
25. University of Helsinki, Helsinki, Finland
26. Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, Australia
27. Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore
28. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
29. Leiden University Medical Center, Leiden, The Netherlands
30. Heart Center Leipzig, Universität Leipzig, Leipzig, Germany
31. Unit of Animal Genomics, WELBIO, GIGA-R & Faculty of Veterinary Medicine, University of Liege, 1 Avenue de l'Hôpital, Liège 4000, Belgium
32. Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom
33. Department of Clinical Physiology, Tampere University Hospital and Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
34. Department of Computer Science, Johns Hopkins University, Baltimore, United States of America
35. Genetics and Genomic Science Department, Icahn School of Medicine at Mount Sinai, New York, United States of America
36. IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany
37. Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany
38. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
39. Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Kanagawa 230-0045, Japan
40. DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany
41. Institute of Clinical Chemistry and Laboratory Medicine, Greifswald University Hospital, Greifswald, Germany
42. Faculty of Genes, Behavior and Health, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
43. Stanford University, Stanford, United States of America
44. Turku University Hospital and University of Turku, Turku, Finland
45. Department of Internal Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands
46. Department of Medicine, Universität Leipzig, Leipzig, Germany
47. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
48. Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands
49. Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
50. University Medical Center Utrecht, Utrecht, The Netherlands
51. Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
52. Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
53. Statistical and Translational Genetics, University of Helsinki, Helsinki, Finland
54. National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland
55. Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
56. Institute of Human Genetics, Technical University Munich, Munich, Germany.
57. Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States of America
58. Department of Medicine, University of Washington, Seattle, United States of America
59. Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton Cambridge, United Kingdom
60. Genome Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
61. Division of Computational Genomics and Systems Genetics, German Cancer Research Center, 69120 Heidelberg, Germany
62. Departments of Biomedical Engineering and Computer Science, Johns Hopkins University, Baltimore, United States of America
63. Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China

Correspondence can be addressed to
Urmo Võsa (urmo.vosa@gmail.com)
Annique Claringbould (anniqueclaringbould@gmail.com)
Lude Franke (ludefranke@gmail.com)

Summary

While many disease-associated variants have been identified through genome-wide association studies, their downstream molecular consequences remain unclear.

To identify these effects, we performed *cis*- and *trans*-expression quantitative trait locus (eQTL) analysis in blood from 31,684 individuals through the eQTLGen Consortium.

We observed that *cis*-eQTLs can be detected for 88% of the studied genes, but that they have a different genetic architecture compared to disease-associated variants, limiting our ability to use *cis*-eQTLs to pinpoint causal genes within susceptibility loci.

In contrast, *trans*-eQTLs (detected for 37% of 10,317 studied trait-associated variants) were more informative. Multiple unlinked variants, associated to the same complex trait, often converged on *trans*-genes that are known to play central roles in disease etiology.

We observed the same when ascertaining the effect of polygenic scores calculated for 1,263 genome-wide association study (GWAS) traits. Expression levels of 13% of the studied genes correlated with polygenic scores, and many resulting genes are known to drive these traits.

Main text

Expression quantitative trait loci (eQTLs) have become a common tool to interpret the regulatory mechanisms of the variants associated with complex traits through genome-wide association studies (GWAS). C*is*-eQTLs, where gene expression levels are affected by a nearby single nucleotide polymorphism (SNP) (<1 megabases; Mb), in particular, have been widely used for this purpose. However, *cis*-eQTLs from the genome tissue expression project (GTEx) explain only a modest proportion of disease heritability[1].

In contrast, *trans*-eQTLs, where the SNP is located distal to the gene (>5Mb) or on other chromosomes, can provide insight into the effects of a single variant on many genes. *Trans*-eQTLs identified before[1–7] have already been used to identify putative key driver genes that contribute to disease[8]. However, *trans*-eQTL effects are generally much weaker than those of *cis*-eQTLs, requiring a larger sample size for detection.

While *trans*-eQTLs are useful for the identification of the downstream effects of a single variant, a different approach is required to determine the combined consequences of trait-associated variants. Polygenic scores (PGS) have been recently applied to sum genome-wide risk for several diseases and likely will improve clinical care[9,10]. However, the exact consequences of different PGS at the molecular level, and thus the contexts in which a polygenic effects manifest themselves, are largely unknown. Here, we systematically investigate *trans*-eQTLs as well as associations between PGS and gene expression (expression quantitative trait score, eQTS) to determine how genetic effects influence and converge on genes and pathways that are important for complex traits.

To maximize the statistical power to detect eQTL and eQTS effects, we performed a large-scale meta-analysis in 31,684 blood samples from 37 cohorts (assayed using three gene expression

platforms) in the context of the eQTLGen Consortium. This allowed us to identify significant *cis*-eQTLs for 16,989 genes, *trans*-eQTLs for 6,298 genes and eQTS effects for 2,568 genes (**Figure 1A**), revealing complex regulatory effects of trait-associated variants. We combine these results with additional data layers and highlight a number of examples where we leverage this resource to infer novel biological insights into mechanisms of complex traits. We hypothesize that analyses identifying genes further downstream are more cell-type specific and more relevant for understanding disease (**Figure 1B**).

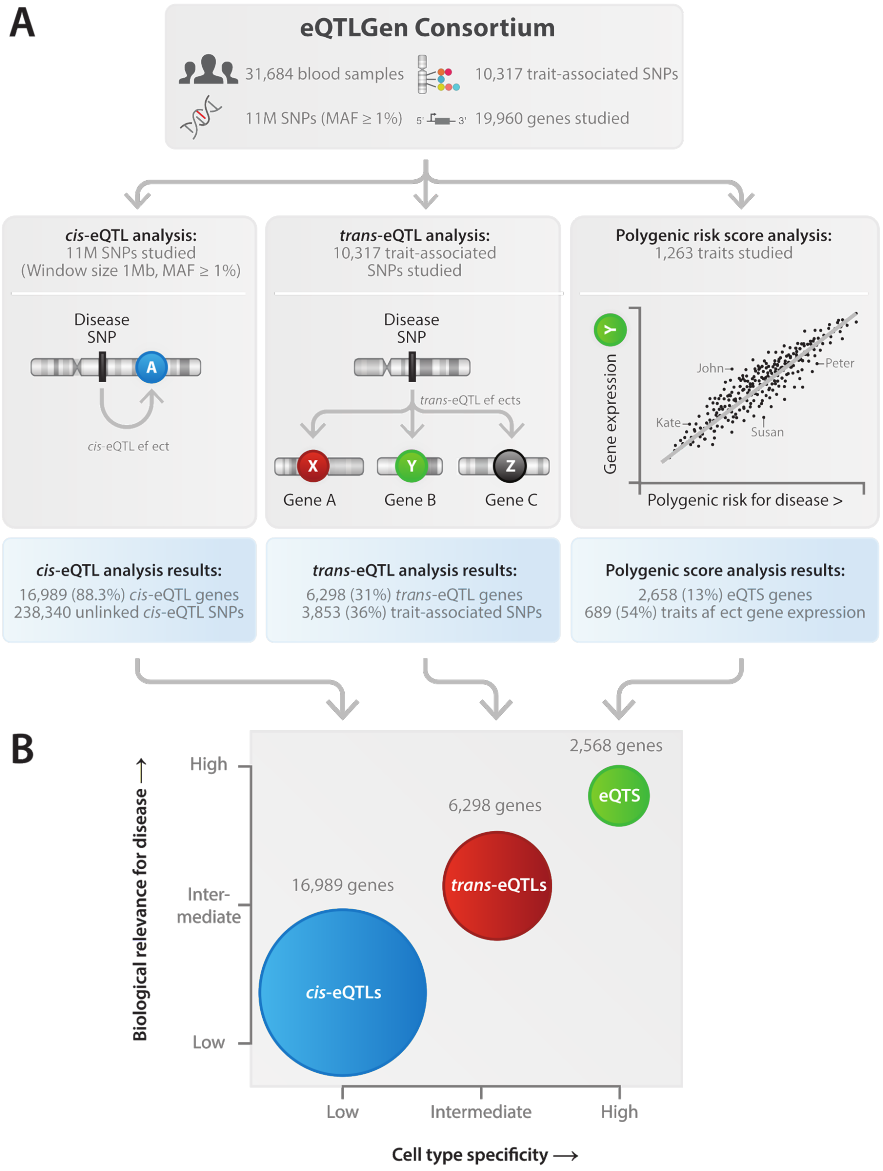


Figure 1. Overview of the study. (**A**) Overview of main analyses and their results. (**B**) Model of genetic effects on gene expression. *Cis*-eQTL are common and widely replicable in different tissues and cell types, whereas *trans*-eQTLs and eQTS are more cell type specific. The biological insight derived from our *cis*-eQTL results are usually not well interpretable in the context of complex traits, suggesting that weaker distal effects give additional insight about biological mechanisms leading to complex traits.

Local genetic effects on gene expression in blood are widespread and replicable in other tissues

Using eQTLGen consortium data from 31,684 individuals, we performed *cis*-eQTL, *trans*-eQTL and eQTS meta-analyses (**Figure 1A**, **Supplementary Table 1**). Different expression profiling platforms were integrated using a data-driven method (**Online Methods**). To ensure the robustness of the identified eQTLs, we performed eQTL discovery per platform and replicated resulting eQTLs in the other platforms, observing excellent replication rates and consistency of allelic directions (**Online Methods**, **Supplementary Note**, **Extended Data Figure 1A-C**). We identified significant *cis*-eQTLs (SNP-gene distance <1Mb, gene-level False Discovery Rate (FDR)<0.05; **Online Methods**) for 16,989 unique genes (88.3% of autosomal genes expressed in blood and tested in *cis*-eQTL analysis; **Figure 1A**). Out of 10,317 trait-associated SNPs tested, 1,568 (15.2%) were in high linkage disequilibrium (LD) with the lead eQTL SNP showing the strongest association for a *cis*-eQTL gene, (R^2>0.8; 1kG p1v3 EUR; **Supplementary Table 2**; **Online Methods**). Genes highly expressed in blood but not under any detectable *cis*-eQTL effect were more likely ($P=2\times10^{-6}$; Wilcoxon two-sided test; **Figure 2A**) to be intolerant to loss-of-function mutations in their coding region[11], suggesting that eQTLs on such gene would interfere with the normal functioning of the organism.

We observed that 92% of the lead *cis*-eQTL SNPs map within 100kb of the gene (**Figure 2D**), and this increased to 97.2% when only looking at the 20% of the genes with the strongest lead *cis*-eQTL effects. Of these strong *cis*-eQTLs, 84.1% of the lead eQTL SNPs map within 20kb of the gene. GWAS simulations[12] indicate that lead GWAS signals map within 33.5kb from the causal variant in 80% of cases, which suggests that our top SNPs usually tag causal variants that map

directly into either the promoter region, the transcription start site (TSS), the gene body, or the transcription end site (TES). For strong *cis*-eQTLs we observed that lead *cis*-eQTL SNPs located >100kb from the TSS or TES overlap capture Hi-C contacts (37%; **Figure 2E**) more often than short-range *cis*-eQTL effects (16%; Chi^2 test $P = 2\times10^{-5}$), indicating that, for long-range *cis*-eQTLs the SNP and gene often physically interact to cause the *cis*-eQTL effect. For instance, a capture Hi-C contact for *IRS1* overlapped the lead eQTL SNP, mapping 630kb downstream from *IRS1* (**Figure 2F**).

We observed that our sample-size improved fine-mapping: for 5,440 protein-coding *cis*-eQTL genes that we had previously identified in 5,311 samples[1] we now observe that the lead SNP typically map closer to the *cis*-eQTL gene (**Extended Data Figure 4**).

Cis-eQTLs showed directional consistency across tissues: in 47 postmortem tissues (GTEx v7[13]) we observed an average of 14.8% replication rate (replication FDR<0.05 in GTEx; median 15.1%; range 3.6-29.7%; whole blood tissue excluded) and on average a 95.0% concordance in allelic directions (median 95.3%, range 86.7-99.3%; whole blood tissue excluded) among the *cis*-eQTLs that significantly replicated in GTEx (**Extended Data Figure 5**, **Supplementary Note** and **Supplementary Table 3**).

However, our lead *cis*-eQTL SNPs show significantly different epigenetic histone mark characteristics, as compared to 3,668 SNPs identified in GWAS (and associated to blood related traits or immune-mediated diseases to minimize potential confounding). We observed significant differences for 20 out of 32 tested histone marks with H3K36me3, H3K27me3, H3K79me1 and H2BK20ac showing the strongest difference (Wilcoxon $P = 10^{-39}$, 10^{-21}, 10^{-19} and 10^{-18},

respectively), suggesting that *cis*-eQTLs have a different genetic architecture, as compared to complex traits and diseases.

We tested this for 16 well-powered complex traits (Supplementary Table 20) and observed that genes prioritized by combining *cis*-eQTL and GWAS data using summary statistics based Mendelian randomization (SMR[14]**; Online Methods**) did not overlap significantly more with genes prioritized through an alternative method (DEPICT) that does not use any *cis*-eQTL information[15]. While the genes prioritized with SMR were informative, and enriched for relevant pathways for several immune traits (**Supplementary Table 20**)**,** non-blood-trait-prioritized genes were difficult to interpret in the context of disease. Moreover, the lack of enriched overlap between DEPICT and SMR indicates that employing *cis*-eQTL information does not necessarily clarify which genes are causal for a given susceptibility locus. As such, some caution is warranted when using a single *cis*-eQTL repository for interpretation of GWAS.

One third of trait-associated variants have *trans*-eQTL effects

An alternative strategy for gaining insight into the molecular functional consequences of disease-associated genetic variants is to ascertain *trans*-eQTL effects. We tested 10,317 trait-associated SNPs ($P \leq 5\times10^{-8}$; **Online Methods**, **Supplementary Table 2**) for *trans*-eQTL effects (SNP-gene distance >5Mb, FDR < 0.05) to better understand their downstream consequences. We identified a total of 59,786 significant *trans*-eQTLs (FDR<0.05; **Supplementary Table 4**, **Extended Data Figure 6**), representing 3,853 unique SNPs (37% of tested GWAS SNPs) and 6,298 unique genes (32% of tested genes; **Figure 1A**). When compared to the previous largest *trans*-eQTL meta-analysis[1] (N=5,311; 8% of trait-associated SNPs with a significant *trans*-eQTL), these results

indicate that a large sample size is critical for identifying downstream effects. Colocalization analyses in a subset of samples (n=4,339; **Supplementary Note**) using COLOC[16] estimated that 52% of *trans*-eQTL signals colocalize with at least one *cis*-eQTL signal (posterior probability > 0.8; **Extended Data Figure 7A-B**). Corresponding colocalizing *cis*-eQTL genes were enriched for transcription factor activity ("regulation of transcription from RNA polymerase II promoter"; $P < 1.3\times10^{-9}$; **Extended Data Figure 7C**). Finally, highly expressed genes without a detectable *trans*-eQTL effect were more likely to be intolerant to loss-of-function variants ($P=6.4\times10^{-7}$; Wilcoxon test, **Figure 2B**), similar to what we observed for *cis*-eQTLs.

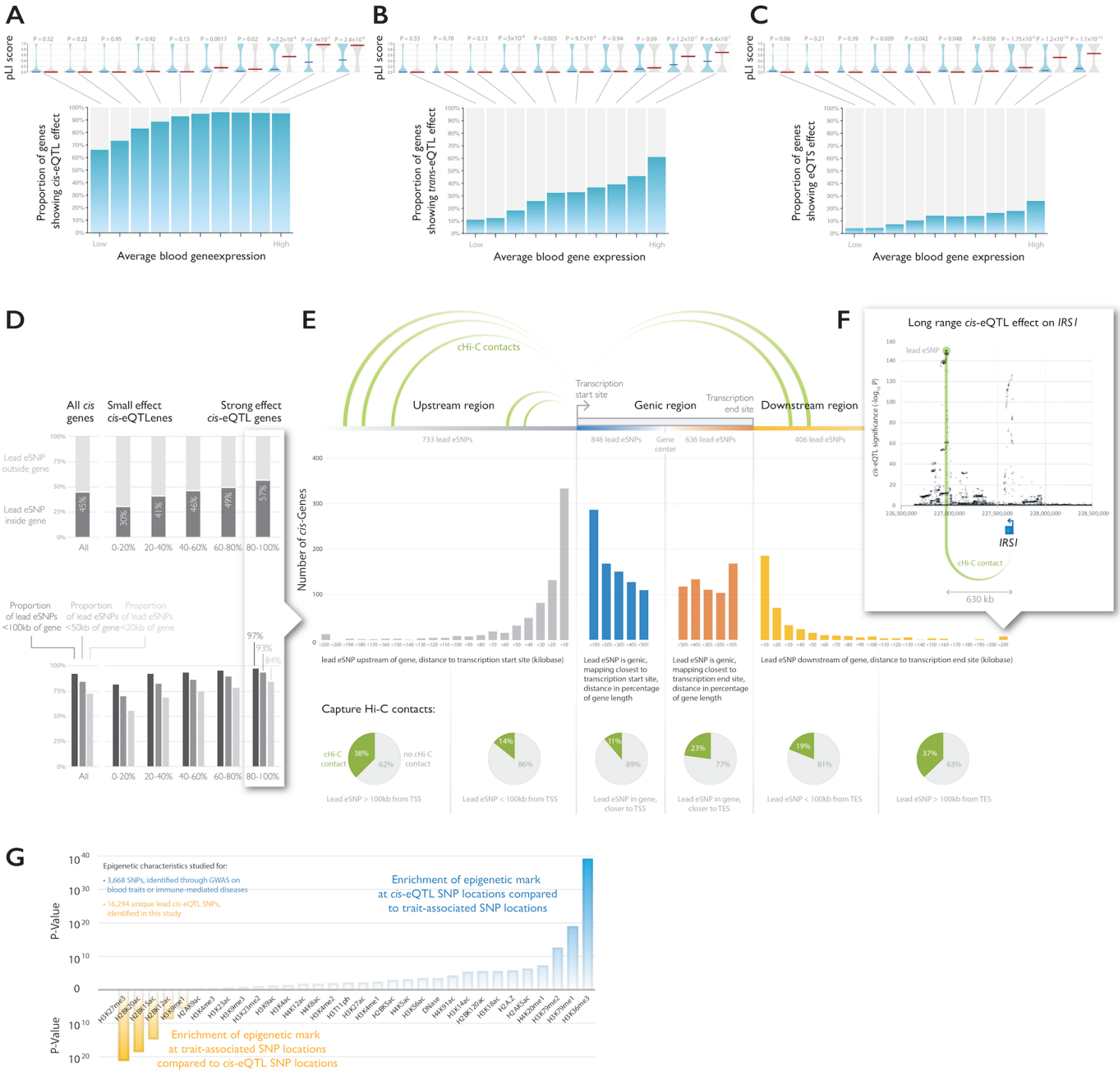


Figure 2. Results of the *cis*- and *trans*-eQTL analysis. All genes tested in (**A**) *cis*-eQTL analysis, (**B**) *trans*-eQTL analysis, and (**C**) eQTS analysis were divided into 10 bins based on average expression levels of the genes in blood. Highly expressed genes without any eQTL effect (grey bars) were less tolerant to loss-of-function variants (Wilcoxon test on pLI scores). Indicated are medians per bin. (**D**) Genes with strong effect sizes are more likely to have a lead SNP within (top panel) or close to the gene (bottom panel) (**E**) Top *cis*-eQTL SNPs positioning further from transcription start site (TSS) and transcription end site (TES) are more likely to overlap capture Hi-C contacts with TSS. (**F**) Enrichment analyses on epigenetic marks of *cis*-eQTL lead SNPs, compared to SNPs identified through GWAS and associated to blood-related or immune-mediated diseases, reveal significant differences in epigenetic characteristics.

In order to study the biological nature of the *trans*-eQTLs we identified, we conducted several enrichment analyses (**Supplementary Note, Extended Data Figure 8, Figure 3**). We observed 2.2 fold enrichment for known transcription factor (TF) - target gene pairs[17] (Fisher's exact test P = 10^{-62}; **Supplementary Note**), with the fold enrichment increasing to 3.2 (Fisher's exact test P < 10^{-300}) when co-expressed genes were included to TF targets. Those genes are potentially further downstream of respective TF targets in the molecular network. Similarly, we observed 1.19 fold enrichment of protein-protein interactions[18] among *trans*-eQTL gene-gene pairs (Fisher's exact test P=0.05). Some of these *cis-trans* gene pairs encode subunits of the same protein complex (e.g. *POLR3H* and *POLR1C*). While significant *cis-trans* gene pairs were enriched for gene pairs showing co-expression (Pearson R > 0.4; Fisher's exact test P=10^{-35}), we did not observe any enrichment of chromatin-chromatin contacts[19] (0.99 fold enrichment; Fisher's exact test P=0.3). Using the subset of 3,831 samples from BIOS, we also ascertained whether the *trans*-eQTL effect was mediated through a gene that mapped within 100kb from the *trans*-eSNP (i.e. using the *cis*-gene as G × E term). We observed significant interaction effects for 523 SNP-*cis*-*trans*-gene

combinations (FDR < 0.05; **Supplementary Table 5**), reflecting a 5.3 fold enrichment compared to what is expected by chance (Fisher's exact $P = 7 \times 10^{-67}$). For instance, for rs7045087 (associated to red blood cell counts) we observed that the expression of interferon gene *DDX58* (mapping 38bp downstream from rs7045087) significantly interacted with *trans*-eQTL effects on interferon genes *HERC5, OAS1, OAS3, MX1, IFIT1, IFIT2, IFIT5, IFI44, IFI44L, RSAD2* and *SAMD9* (**Extended Data Figure 9**).

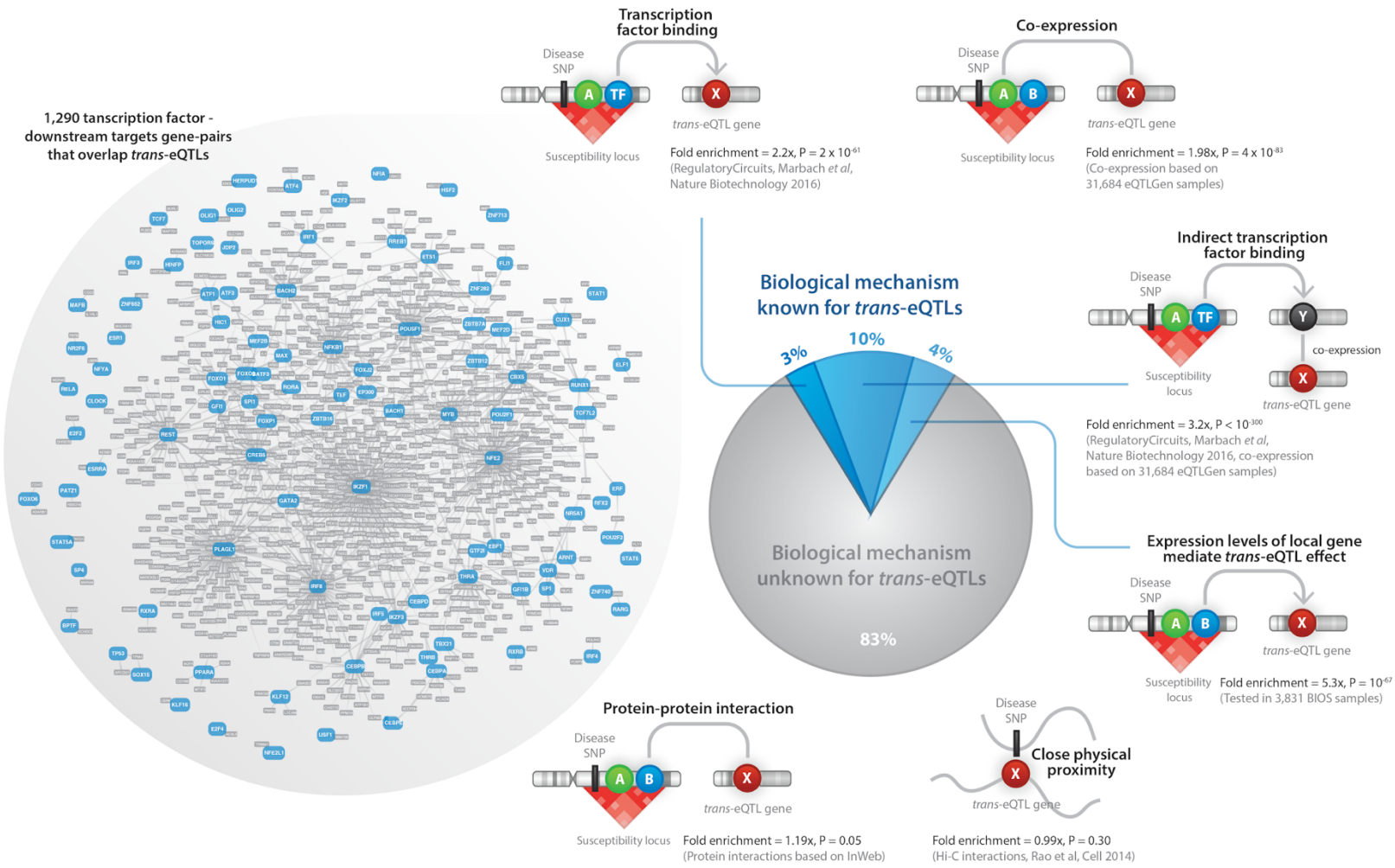


Figure 3. Mechanisms leading to *trans*-eQTLs. Shown are the results of enrichment analyses for known TF associations, HiC contacts, protein-protein interactions, gene co-expression and mediation analyses.

We estimate that 17.4% of the identified *trans*-eQTLs are explainable by (indirect) TF binding or mediation by *cis*-genes (**Supplementary Note**). This leaves 82.6% of the observed *trans*-eQTL effects unexplained. While it is likely that many of these *trans*-eQTLs reflect unknown (indirect) effects of TFs, we speculate that novel and unknown regulatory mechanisms could also play a role. By making all *trans*-eQTL results (irrespective of their statistical significance) publicly available, we envision this dataset will help to yield such insight in the future.

To estimate the proportion of loci where the trait-associated SNP explained the *trans*-eQTL signal in the locus, we performed locus-wide conditional *trans*-eQTL analysis in a subset of 4,339 samples for 12,991 *trans*-eQTL loci (**Online Methods; Extended Data Figure 10; Supplementary Table 6**). In 43% of these loci, we observed that the trait-associated SNP was in high LD with the *trans*-eQTL SNP having the strongest association in the locus ($R^2 > 0.8$, 1kG p1v3 EUR; **Supplementary Table 7**). For 95 cases, the strongest *cis*- and *trans*-eQTL SNPs were both in high LD with GWAS SNP ($R^2 > 0.8$ between top SNPs, 1kG p1v3 EUR; **Supplementary Table 7**).

The majority (64%) of *trans*-eQTL SNPs have previously been associated with blood composition phenotypes, such as platelet count, white blood cell count and mean corpuscular volume[20]. In comparison, blood cell composition SNPs from the same study comprised only 20.7% od all the tested trait-associated SNPs. This was expected, since SNPs that regulate the abundance of a specific blood cell type would result in *trans*-eQTL effects on genes, specifically expressed in that cell type.

Therefore, we aimed to distinguish *trans*-eQTLs caused by intracellular molecular mechanisms from blood cell type QTLs using eQTL data from lymphoblastoid cell line (LCL), induced pluripotent cells (iPSCs), several purified blood cell types (CD4+, CD8+, CD14+, CD15+, CD19+, monocytes and platelets) and blood DNA methylation QTL data. In total, 3,853 (6.4%) of *trans*-eQTLs showed significant replication in at least one cell type or in the methylation data (**Extended Data Figure 11**, **Supplementary Table 11A**). While this set of *trans*-eQTLs (denoted as the “intracellular eQTLs”) is less likely to be driven by cell type composition, we acknowledge that the limited sample size of the available *trans*-eQTL replication datasets make our replication effort very conservative. Furthermore, *trans*-eQTLs caused by variants associated with cell type proportions may be

informative for understanding the biology of a trait. Therefore, we did not remove these kinds of *trans*-eQTLs from our interpretative analyses.

Next, we aimed to replicate the identified *trans*-eQTLs in the tissues from GTEx[13]. Although the replication rate was very low (0-0.03% of *trans*-eQTLs replicated in non-blood tissues, FDR < 0.05, same allelic direction; **Supplementary Table 11B**), we did observe an inflation of signal (median chi-squared statistic) for identified *trans*-eQTLs in several GTEx tissues (**Extended Data Figure 12**). Non-blood tissues showing the strongest inflation were liver, heart atrial appendage and non-sun-exposed skin.

Trans-eQTLs are effective for discerning the genetic basis of complex traits

As described above, *trans*-eQTLs can arise due to *cis*-eQTL effects on TFs, whose target genes show *trans*-eQTL effects. We describe below such examples, but also highlight *trans*-eQTLs where the eQTL SNP works through a different mechanism. **Combining *cis*- and *trans*-eQTL effects can pinpoint the genes acting as drivers of *trans*-eQTL effects.** For example, the age-of-menarche-associated SNP rs1532331[21] is in high LD with the top *cis*-eQTL effect for transcription factor *ZNF131* (R^2 > 0.8, 1kG p1v3 EUR). *Cis*-eQTL and *trans*-eQTL effects for this locus co-localized for 25 out of the 75 downstream genes (**Figure 4A**). In a recent short hairpin RNA knockdown experiment of *ZNF131*[22], three separate cell isolates showed downregulation of four genes that we identified as *trans*-eQTL genes: *HAUS5, TMEM237, MIF4GD* and *AASDH* (**Figure 4A**). *ZNF131* has been hypothesized to inhibit estrogen signaling[23], which may explain how the SNP in this locus contributes to altering the age of menarche.

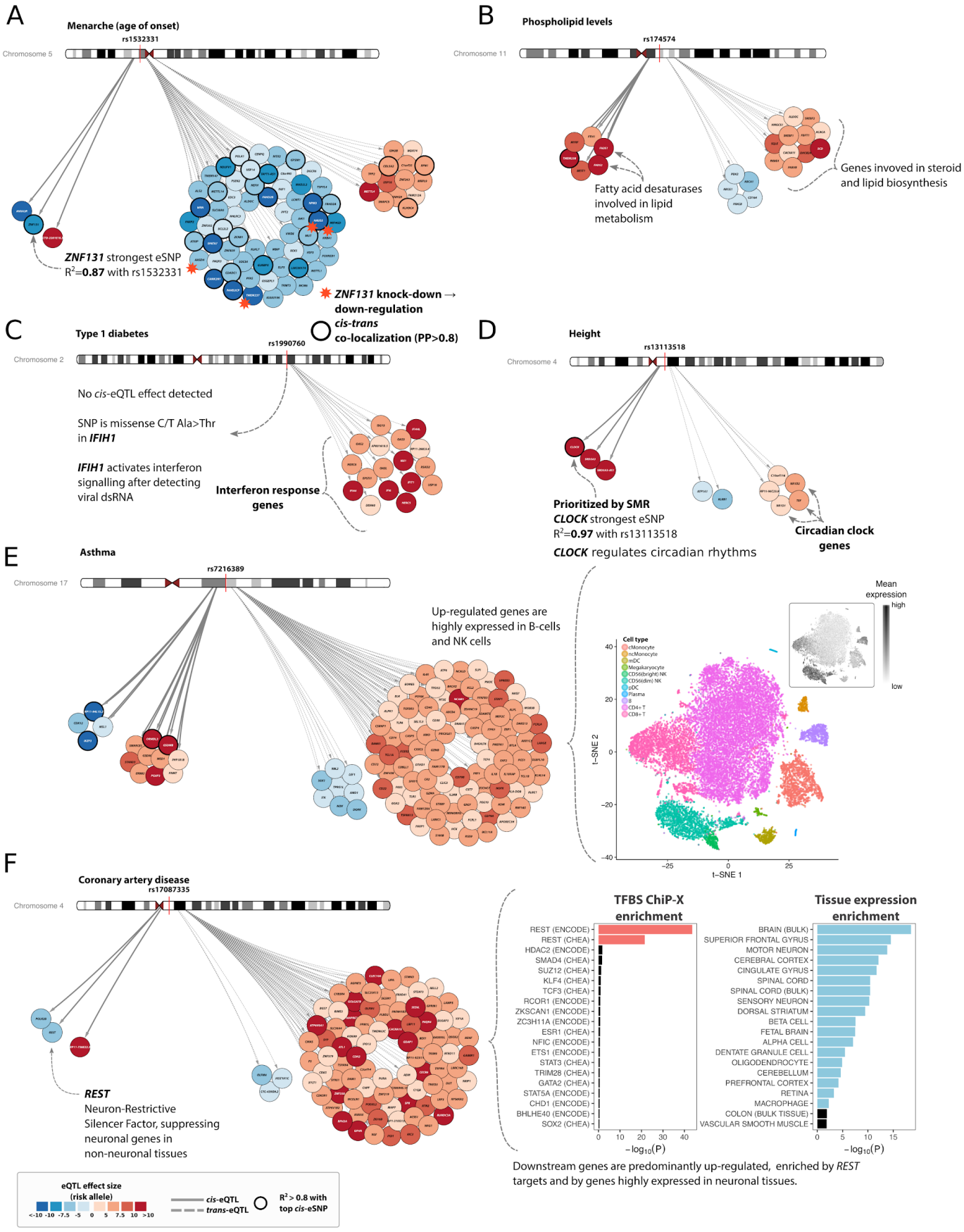


Figure 4. Examples of *cis*- and *trans*-eQTLs. (**A**) *Cis*-eQTL on *ZNF131* is prioritized because several *trans*-eQTL genes are down-regulated by *ZNF131* in functional study. (**B**) Phospholipid-associated SNP shows *cis*-and *trans*-eQTLs on lipid metabolism genes. (**C**) Type I diabetes associated SNP has no *cis*-eQTLs, but *trans*-eQTL genes point to interferon signaling pathway. (**D**) Circadian rhythm genes *CLOCK* (in *cis*) and *NR1D1*, *NR1D2*, *TEF* (in *trans*) identified for height associated SNP. (**E**) eQTLs for asthma SNP tag cell type abundance of B and NK cells. (**F**) *Trans*-eQTL genes for *REST* locus are highly enriched for REST transcription factor targets and for neuronal expression.

***Trans*-eQTLs extend insight for loci with multiple *cis*-eQTL effects.** In the *FADS1*/*FADS2* locus, rs174574 is associated with lipid levels[24] and affects 17 genes in *trans* (**Figure 4B**). The strongest *cis*-eQTLs modulate the expression of *FADS1*, *FADS2* and *TMEM258*, with latter being in high LD with GWAS SNP (R^2>0.8, 1kG p1v3 EUR). *FADS1* and *FADS2* have been implicated[24] since they regulate fatty acid synthesis, and consistent with their function, *trans*-eQTL genes from this locus are highly enriched for triglyceride metabolism ($P < 4.1\times10^{-9}$, GeneNetwork[25] REACTOME pathway enrichment). Since this locus has extensive LD, variant and gene prioritization is difficult: conditional analyses in 4,339 sample subset showed that each of *cis*-eQTL gene is influenced by more than one SNP, but none of these are in high LD with rs174574 (R^2 < 0.8, 1kG p1v3, EUR). As such, our *trans*-eQTL analysis results are informative for implicating *FADS1* and *FADS2*, whereas *cis*-eQTLs are not.

***Trans*-eQTLs can shed light on loci with no detectable *cis*-eQTLs.** rs1990760 is associated with multiple immune-related traits (Type 1 Diabetes (T1D), Inflammatory bowel disease (IBD), Systemic Lupus Erythematosis (SLE) and psoriasis[26–29]). For this SNP we identified 17 *trans*-eQTL effects, but no detectable gene-level *cis*-eQTLs in blood (**Figure 4C**) and GTEx. However, the risk

allele for this SNP causes an Ala946Thr amino acid change in the RIG-1 regulatory domain of MDA5 (encoded by *IFIH1* - Interferon Induced With Helicase C Domain 1), outlining one possible mechanism leading to the observed *trans*-eQTLs. MDA5 acts as a sensor for viral double-stranded RNA, activating interferon I signalling among other antiviral responses. All the *trans*-eQTL genes were up-regulated relative to risk allele to T1D, and 9 (52%) are known to be involved in interferon signaling (**Supplementary Table 12**).

***Trans*-eQTLs can reveal cell type composition effects of the trait-associated SNP.** *Trans*-eQTL effects can also show up as a consequence of a SNP that alters cell-type composition. For example, the asthma-associated SNP rs7216389[30] has 14 *cis*-eQTL effects, most notably on *IKZF3*, *GSDMB*, and *ORMDL3* (**Figure 4E**). SMR prioritized all three *cis*-genes equally (**Extended Data Figure 13**), making it difficult to draw biological conclusions (similar as we observed for the *FADS* locus). However, 94 out of the 104 *trans*-eQTL genes were up-regulated by the risk allele for rs7216389 and were mostly expressed in B cells and natural killer cells[31] (**Figure 4E**). *IKZF3* is part of the Ikaros transcription factor family that regulates B-cell proliferation[31,32], suggesting that a decrease of *IKZF3* leads to an increased number of B cells and concurrent *trans*-eQTL effects caused by cell-type composition differences.

Some *trans*-eQTLs influence genes strongly expressed in tissues other than blood. We observed *trans*-eQTL effects on genes that are hardly expressed in blood, indicating that our *trans*-eQTL effects are informative for non-blood related traits as well: rs17087335, which is associated with coronary artery disease[33], affects the expression of 88 genes in *trans* (**Figure 4F**), that are highly expressed in brain (hypergeometric test, ARCHS4 database, q-value = 2.58×10^{-17}; **Figure 4F, Supplementary Table 13**), but show very low expression in blood. SNPs linked with rs17087335 (R^2 > 0.8, 1kG p1v3 EUR) are associated with height (rs2227901, rs3733309 and

rs17081935)[34,35], and platelet count (rs7665147)[20]. The minor alleles of these SNPs downregulate the nearby gene *REST* (RE-1 silencing transcription factor), although none of these variants is in LD (R^2<0.2, 1kG p1v3 EUR) with the lead *cis*-eQTL SNP for *REST*. *REST* is a TF that downregulates the expression of neuronal genes in non-neuronal tissues[36,37]. It also regulates the differentiation of vascular smooth muscles, and is thereby associated with coronary phenotypes[38]. 85 out of 88 (96.6%) of the *trans*-eQTL genes were upregulated relative to the minor allele and were strongly enriched by transcription factor targets of REST (hypergeometric test for ENCODE REST ChIP-seq, q-value = 1.36×10^{-42} , **Figure 4F**). As such, *trans*-eQTL effects on neuronal genes implicate *REST* as the causal gene in this locus.

***Trans*-eQTLs identify pathways not previously associated with a phenotype.** Some *trans*-eQTLs suggest the involvement of pathways which are not previously thought to play a role for certain complex traits: SMR analysis prioritized *CLOCK* as a potential causal gene in the height-associated locus on chr 4q12 ($P_{SMR}=3\times10^{-25}$; $P_{HEIDI}=0.02$; **Figure 4D**). In line with that, height-associated SNP rs13113518[34] is also in high LD (R^2>0.8, 1kG p1v3 EUR) with the top *cis*-eQTL SNP for *CLOCK*. The upregulated TF CLOCK forms a heterodimer with TF BMAL1, and the resulting protein complex regulates circadian rhythm[39]. Three known circadian rhythm *trans*-eQTL genes (*TEF, NR1D1* and *NR1D2*) showed increased expression for the trait-increasing allele, suggesting a possible mechanism for the observed *trans*-eQTLs through binding of CLOCK:BMAL1. *TEF* is a D-box binding TF whose gene expression in liver and kidney is dependent on the core circadian oscillator and it regulates amino acid metabolism, fatty acid metabolism and xenobiotic detoxification (Gachon et al., 2006)**.** *NR1D1* and *NR1D2* encode the transcriptional repressors Rev-ErbA alpha and beta, respectively, and form a negative feedback loop to suppress *BMAL1* expression[40]. *NR1D1* and *NR1D2* have been reported to be associated

with osteoblast and osteoclast functions[41], revealing a possible link between circadian clock genes and height.

Unlinked trait-associated SNPs converge on the same downstream genes in *trans*. We subsequently ascertained, per trait, whether unlinked trait-associated variants showed *trans*-eQTL effects on the same downstream gene. Here we observed 47 different traits where at least four independent variants affected the same gene in *trans*, 3.4× higher than expected by chance (P = 0.001; two-tailed two-sample test of equal proportions; **Supplementary Table 8**). For SLE, for example, we observed that the gene expression levels of *IFI44L*, *HERC5*, *IFI6*, *IFI44*, *RSAD2*, *MX1*, *ISG15*, *ANKRD55*, *OAS3*, *OAS2*, *OASL* and *EPSTI1* (nearly all interferon genes) were affected by at least three SLE-associated genetic variants, clearly showing the involvement of interferon signaling in SLE (**Figure 5**).

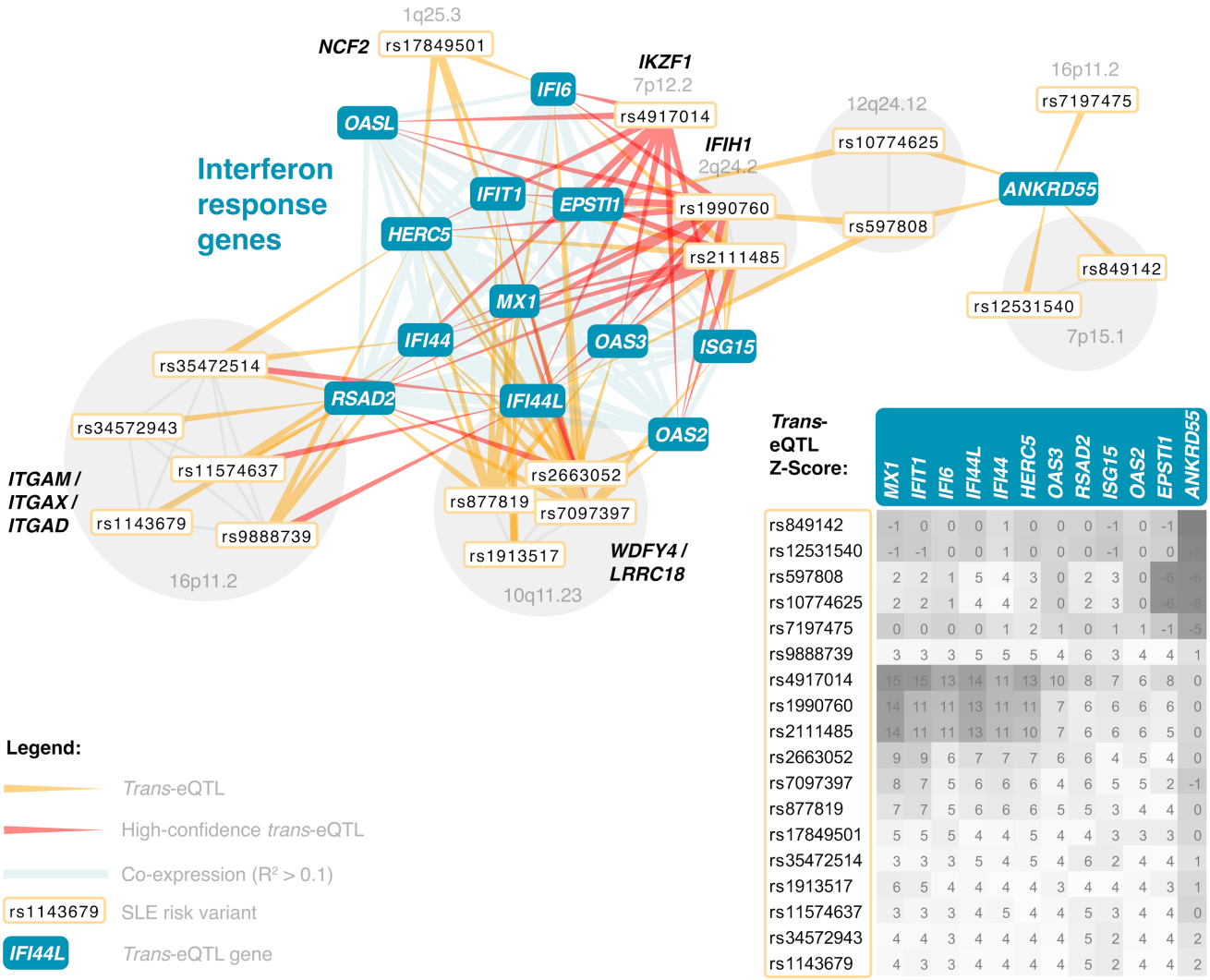


Figure 5. SNPs associated with SLE converge to the shared cluster of interferon response genes. Shown are genes which are affected by at least three independent GWAS SNPs. SNPs in partial LD are grouped together. Heat map indicates the direction and strength of individual *trans*-eQTL effects (Z-scores).

This convergence of multiple SNPs on the same genes lends credence to recent hypotheses with regards to the 'omnigenic' architecture of complex traits[8]: indeed multiple unlinked variants do affect the same 'core' genes. The recent omnigenic model[42] proposes a strategy to partition between core genes, which have direct effects on a disease, and peripheral genes, which can only affect disease risk indirectly through regulation of core genes. In **Supplementary Equations**, we

show that this model also implies a correlation between polygenic risk scores and expression of core genes. We therefore studied this systematically by aggregating multiple associated variants into polygenic scores and ascertaining how they correlate with gene expression levels.

eQTSs identify key driver genes for polygenic traits

To ascertain the coordinated effects of trait-associated variants on gene expression, we used available GWAS summary statistics to calculate PGSs for 1,263 traits in 28,158 samples (**Online Methods, Supplementary Table 14**). We reasoned that when a gene shows expression levels that significantly correlate with the PGS for a specific trait (an expression quantitative trait score; eQTS), the downstream *trans*-eQTL effects of the individual risk variants converge on that gene, and hence, that the gene may be a driver of the disease.

Our meta-analysis identified 18,210 eQTS effects (FDR < 0.05), representing 689 unique traits (54%) and 2,568 unique genes (13%; **Supplementary Table 15**, **Figure 1A**). As expected, most eQTS associations represent blood cell traits (**Extended Data Figure 14, Supplementary Table 16**): for instance the PGS for mean corpuscular volume correlated positively with the expression levels of genes specifically expressed in erythrocytes, such as genes coding for hemoglobin subunits. However, we also identified eQTS associations for genes that are known drivers of other traits.

For example, 11 out of 26 genes associating with the PGS for high density lipoprotein levels (HDL[43,44]; FDR<0.05; **Figure 6A**) have previously been associated with lipid or cholesterol metabolism (**Supplementary Table 18**). *ABCA1* and *ABCG1*, which positively correlated with the PGS for high HDL, mediate the efflux of cholesterol from macrophage foam cells and participate in

HDL formation. In macrophages, the downregulation of both *ABCA1* and *ABCG1* reduced reverse cholesterol transport into the liver by HDL[45] (**Figure 6B**). The genetic risk for high HDL was also negatively correlated with the expression of the low density lipoprotein receptor *LDLR* (strongest eQTS $P=3.35\times10^{-20}$) known to cause hypercholesterolemia[46]. Similarly, the gene encoding the TF SREBP-2, which is known to increase the expression of *LDLR*, was downregulated (strongest eQTS $P=3.08\times10^{-7}$). The negative correlation between *SREBF2* expression and measured HDL levels has been described before[47], indicating that the eQTS reflects an association with the actual phenotype. Zhernakova et al. proposed a model where down-regulation of *SREBF2* results in the effect on its target gene *FADS2*. We did not observe a significant HDL eQTS effect on *FADS2* (all eQTS $P>0.07$), possibly because the indirect effect is too small to detect. We hypothesize that HDL levels in blood can result in a stronger reverse cholesterol transport into the liver, which may result in downregulation of *LDLR*[48]

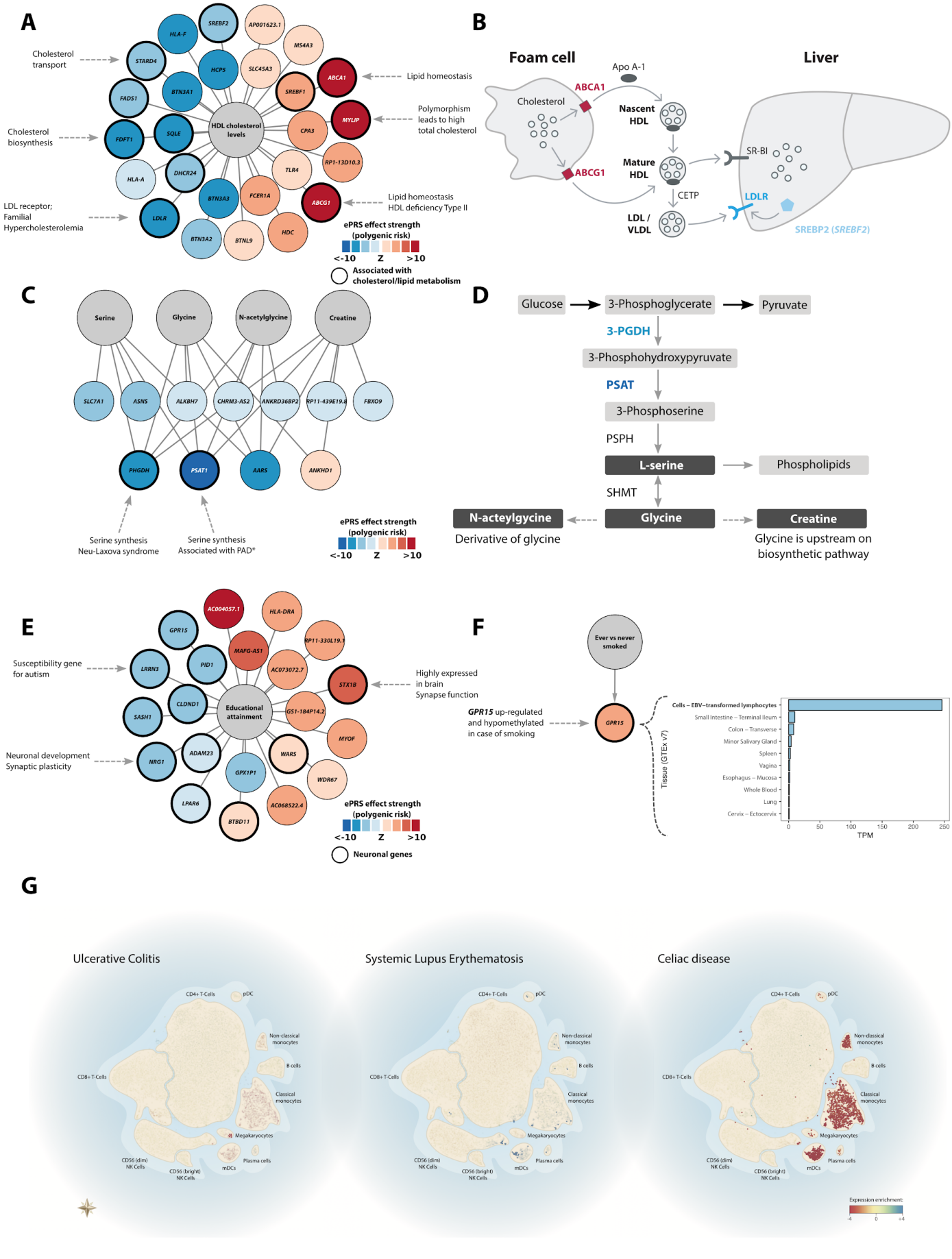


Figure 6. Examples of eQTS. (**A**) Polygenic risk score (PRS) for high density lipoprotein associates to lipid metabolism genes. (**B**) The role of *ABCA1*, *ABCG1*, and *SREBF2* in cholesterol transport. (**C**) Polygenic scores for serine, glycine, n-acetylglycine and creatine levels negatively associate with gene expression of *PHGDH*, *PSAT1*, and *AARS*. (**D**) Serine biosynthesis pathway. (**E**) PRS for educational attainment identifies genes with neuronal functions. (**F**) Polygenic score for smoking status upregulates *GPR15*, which plays a role in lymphocyte differentiation. (**G**) eQTS genes for immune-related diseases are enriched for genes specifically expressed in certain blood cell types.

eQTS analysis also identified genes relevant for non-blood traits, such as the association of *GPR15* ($P=3.7\times10^{-8}$, FDR<0.05; **Figure 6F**) with the trait 'ever versus never smoking'[49]. *GPR15* is a biomarker for smoking[50] that is overexpressed and hypomethylated in smokers[51]. We observe strong *GPR15* expression in lymphocytes (**Figure 6F**), suggesting that the association with smoking could originate from a change in the proportion of T cells in blood[52]. As *GPR15* is involved in T cell homing and has been linked to colitis and inflammatory phenotypes, it is hypothesized to play a key role in smoking-related health risks[53].

The PGS for another non-blood trait, educational attainment[54], correlated significantly with the expression of 21 genes (FDR<0.05; **Figure 6E**, **Supplementary Table 15**). Several of the strongly associated genes are known to be involved in neuronal processes (**Supplementary Table 19**) and show expression in neuronal tissues (GTEx v7, **Extended Data Figure 15**). *STX1B* (strongest eQTS $P=1.3\times10^{-20}$) is specifically expressed in brain, and its encoded protein, syntaxin 1B, participates in the exocytosis of synaptic vesicles and synaptic transmission[55]. Another gene highly expressed in brain, *LRRN3* (Leucine-rich repeat neuronal protein 3; strongest eQTS $P=1.7\times10^{-11}$) was negatively associated with the PGS for educational attainment, and has been associated with autism susceptibility[56]. The downregulated *NRG1* (neuregulin 1; strongest eQTS $P=4.5\times10^{-7}$),

encodes a well-established growth factor involved in neuronal development and has been associated to synaptic plasticity[57]. *NRG1* was also positively associated with the PGS for monocyte levels[20] (strongest eQTS $P=1.5\times10^{-7}$), several LDL cholesterol traits (e.g. medium LDL particles[44]; strongest eQTS $P=6.2\times10^{-8}$), coronary artery disease[33] (strongest eQTS $P=1.5\times10^{-6}$) and body mass index in females[58] (strongest eQTS $P=9.2\times10^{-12}$).

eQTS can also identify pathways known to be associated with monogenic diseases**.** For example, the PGSs for serine, glycine, the glycine derivative n-acetylglycine and creatine[59,60] (**Figure 6C**) were all negatively associated with the gene expression levels of *PHGDH*, *PSAT1* and *AARS* ($P < 5.3\times10^{-7}$). *PHGDH* and *PSAT1* encode crucial enzymes that regulate the synthesis of serine and, in turn, glycine[61] (**Figure 6D**), while n-acetylglycine and creatine form downstream of glycine[62]. Mutations in *PSAT1* and *PHGDH* can result in serine biosynthesis defects including phosphoserine aminotransferase deficiency[63], phosphoglycerate dehydrogenase deficiency[64], and Neu-Laxova syndrome[65], all diseases characterized by low concentrations of serine and glycine in blood and severe neuronal manifestations. *AARS* encodes alanyl-tRNA synthetase, which links alanine to tRNA molecules. A mutation in *AARS* has been linked to Charcot Marie Tooth disease[66], while the phenotypically similar hereditary sensory neuropathy type 1 (HSN1[67]) can be caused by a mutation in the gene encoding serine palmitoyltransferase. The gene facilitates serine's role in sphingolipid metabolism[68]. Disturbances in this pathway are hypothesized to be central in the development the neuronal symptoms[69], suggesting a link between *AARS* expression and the serine pathway. Unexpectedly, the genetic risk for higher levels of these amino acids was associated with lower expression of *PHGDH*, *PSAT1*, and *AARS*, implying the presence of a negative feedback loop that controls serine synthesis.

We next evaluated 6 immune diseases for which sharing of loci has been reported previously, and also observed sharing of downstream eQTS effects for these diseases (**Supplementary Table 20**). For example, the interferon gene *STAT1* was significantly associated with T1D, celiac disease (CeD), IBD and primary biliary cirrhosis (PBC). However, some of these genes are also marker genes for specific blood-cell types, such as *CD79A*, which showed a significant correlation with T1D and PBC. To test whether disease-specific eQTS gene signatures are reflected by blood cell proportions, we investigated single-cell RNA-seq data[31] (**Online Methods**; **Figure 6G**). For ulcerative colitis (a subtype of IBD), we observed significant depletion of expression in megakaryocytes. SLE eQTS genes were enriched for antigen presentation (GeneNetwork $P=1.3\times10^{-5}$) and interferon signaling (GeneNetwork $P=1.4\times10^{-4}$), consistent with the well-described interferon signature in SLE patients[70,71]. Moreover, the SLE genes were significantly enriched for expression in mature dendritic cells, whose maturation depends on interferon signaling[72]. For CeD, we observed strong depletion of eQTS genes in monocytes and dendritic cells, and a slight enrichment in CD4+ and CD8+ T cells. The enrichment of cytokine (GeneNetwork $P=1.6\times10^{-15}$) and interferon (GeneNetwork $P=7.8\times10^{-13}$) signaling among the CeD eQTS genes is expected as a result of increased T cell populations.

Cell-type-specificity of eQTS associations

We next ascertained to what extent these eQTS associations can be replicated in non-blood tissues. We therefore aimed to replicate the significant eQTS effects in 1,460 LCL samples and 762 iPSC samples. Due to the fact these cohorts have a comparatively low sample sizes and study different cell types, we observed limited replication: 10 eQTS showed significant replication effect (FDR<0.05) in the LCL dataset, with 9 out of those (90%) showing the same effect direction as in

the discovery set (**Extended Data Figure 16A, Supplementary Table 17**). For iPSCs, only 5 eQTS showed a significant effect (**Extended Data Figure 16B, Supplementary Table 17**). Since only a few eQTS associations are significant in non-blood tissues and the majority of identified eQTS associations are for blood-related traits, we speculate these effects are likely to be highly cell-type specific. This indicates that large-scale eQTL meta-analyses in other tissues could uncover more genes on which trait-associated SNPs converge.

Discussion

We here performed *cis*-eQTL, *trans*-eQTL and eQTS analyses in 31,684 blood samples, reflecting a six-fold increase over earlier large-scale studies[1,5]. We identified *cis*-eQTL effects for 88.3% and *trans*-eQTL effects for 32% of all genes that are expressed in blood.

We observed that *cis*-eQTL SNPs map close to the TSS or TES of the *cis*-gene: for the top 20% strongest *cis*-eQTL genes, 84.1% of the lead eQTL SNPs map within 20kb of the gene, indicating that these are variants immediately adjacent to the start or end of transcripts that primarily drive *cis*-eQTL effects. The trait-associated variants that we studied showed a different pattern: 77.4% map within 20kb of the closest protein-coding gene, suggesting that the genetic architecture of *cis*-eQTLs is different from disease-associated variants. This is supported by the epigenetic differences that we observed between these two groups and can also partly explain, why we did not observe significantly increased overlap between genes prioritized using pathway enrichment analysis[15] and genes prioritized using our *cis*-eQTLs.

In contrast, for numerous traits we observed that multiple unlinked *trans*-eQTL variants often converge on genes with a known role in the biology for these traits (e.g. the involvement of interferon genes in SLE).

We therefore focused on *trans*-eQTL and eQTS results to gain insight into trait-relevant genes and pathways (**Figures 4, 6**). We estimate that 17.4% of our *trans*-eQTLs are driven by transcriptional regulation, whereas the remaining fraction is driven by not-yet-identified mechanisms. Our results support a model which postulates that, compared to *cis*-eQTLs, weaker distal and polygenic effects converge on core (key driver) genes that are more relevant to the traits and more specific for trait-relevant cell types (**Figure 1B**). The examples we have highlighted demonstrate how insights can be gained from our resource, and we envision similar interpretation strategies can be applied to the other identified *trans*-eQTL and eQTS effects. The catalog of genetic effects on gene expression we present here (available at www.eqtlgen.org) is a unique compendium for the development and application of novel methods that prioritize causal genes for complex traits[14,73], as well as for interpreting the results of genome-wide association studies.

Methods

Cohorts

eQTLGen Consortium data consists of 31,684 blood and PBMC samples from 37 datasets, pre-processed in a standardized way and analyzed by each cohort analyst using the same settings (**Online Methods**). 26,886 (85%) of the samples added to discovery analysis were whole blood samples and 4,798 (15%) were PBMCs, and the majority of samples were of European ancestry (**Supplementary Table 1**). The gene expression levels of the samples were profiled by Illumina (N=17,421; 55%), Affymetrix U291 (N=2,767; 8.7%), Affymetrix HuEx v1.0 ST (N=5,075; 16%) expression arrays and by RNA-seq (N=6,422; 20.3%). A summary of each dataset is outlined in **Supplementary Table 1.** Detailed cohort descriptions can be found in the **Supplementary Note**. Each of the cohorts completed genotype and expression data pre-processing, PGS calculation, *cis*-eQTL-, *trans*-eQTL- and eQTS-mapping, following the steps outlined in the online analysis plans, specific for each platform (see **URL**s) or with slight alterations as described in **Supplementary Table 1** and the **Supplementary Note**. All but one cohort (Framingham Heart Study), included non-related individuals into the analysis.

Genotype data preprocessing

The primary pre-processing and quality control of genotype data was conducted by each cohort, as specified in the original publications and in the **Supplementary Note**. The majority of cohorts used genotypes imputed to 1kG p1v3 or a newer reference panel. GenotypeHarmonizer[74] was used to harmonize all genotype datasets to match the GIANT 1kG p1v3 ALL reference panel and

to fix potential strand issues for A/T and C/G SNPs. Each cohort tested SNPs with minor allele frequency (MAF) > 0.01, Hardy-Weinberg P-value > 0.0001, call rate > 0.95, and MACH r^2 > 0.5.

Expression data preprocessing

Illumina arrays

Illumina array datasets expression were profiled by HT-12v3, HT-12v4 and HT-12v4 WGDASL arrays. Before analysis, all the probe sequences from the manifest files of those platforms were re-mapped to GRCh37.p10 human genome and transcriptome, using SHRiMP v2.2.3 aligner[75] and allowing 2 mismatches. Probes mapping to multiple locations in the genome were removed from further analyses.

For Illumina arrays, the raw unprocessed expression matrix was exported from GenomeStudio. Before any pre-processing, the first two principal components (PCs) were calculated on the expression data and plotted to identify and exclude outlier samples. The data was normalized in several steps: quantile normalization, $\log_2$ normalization, probe centering and scaling by the equation $Expression_{Probe,Sample}$ = ($Expression_{Probe,Sample}$ - MeanProbe) / $Std.Dev._{Probe}$. Genes showing no variance were removed. Next, the first four multidimensional scaling (MDS) components, calculated based on non-imputed and pruned genotypes using plink v1.07[76], were regressed out of the expression matrix to account for population stratification. We further removed up to 20 first expression-based PCs that were not associated to any SNPs, as these capture non-genetic variation in expression. Each cohort also ran MixupMapper[77] software to identify incorrectly labeled genotype-expression combinations, and to remove identified sample mix-ups.

Affymetrix arrays

Affymetrix array-based datasets used the expression data previously pre-processed and quality controlled as indicated in the **Supplementary Note**.

RNA-seq

Alignment, initial quality control and quantification differed slightly across datasets, as described in the **Supplementary Note**. Each cohort removed outliers as described above, and then used Trimmed Mean of M-values (TMM) normalization and a counts per million (CPM) filter to include genes with >0.5 CPM in at least 1% of the samples. Other steps were identical to Illumina processing, with some exceptions for the BIOS Consortium datasets (**Supplementary Note**).

Cis-eQTL mapping

Cis-eQTL mapping was performed in each cohort using a pipeline described previously[1]. In brief, the pipeline takes a window of 1Mb upstream and 1Mb downstream around each SNP to select genes or expression probes to test, based on the center position of the gene or probe. The associations between these SNP-gene combinations was calculated using a Spearman correlation. Next, 10 permutation rounds were performed by shuffling the links between genotype and expression identifiers and re-calculating associations. The false discovery rate (FDR) was determined using 10 meta-analyzed permutations: for each gene in the real analysis, the most significant association was recorded, and the same was done for each of the permutations,

resulting in a gene-level FDR. *Cis*-eQTLs with a gene-level FDR < 0.05 (corresponding to $P < 1.829\times10^{-5}$) and tested in at least two cohorts were deemed significant.

Trans-eQTL mapping

Trans-eQTL mapping was performed using a previously described pipeline[1] while testing a subset of 10,317 SNPs previously associated with complex traits. We required the distance between the SNP and the center of the gene or probe to be >5Mb. To maximize the power to identify *trans*-eQTL effects, the results of the summary statistics based or iterative conditional *cis*-eQTL mapping analyses (**Supplementary Note**) were used to correct the expression matrices before *trans*-eQTL mapping. For that, top SNPs for significant conditional *cis*-eQTLs were regressed out from the expression matrix. Finally, we removed potential false positive *trans*-eQTLs caused by reads cross-mapping with *cis* regions (**Supplementary Note**).

Genetic risk factor selection

Genetic risk factors were downloaded from three public repositories: the EBI GWAS Catalogue[78] (downloaded 21.11.2016), the NIH GWAS Catalogue and Immunobase (www.immunobase.org; accessed 26.04.2016), applying a significance threshold of $P \leq 5\times10^{-8}$. Additionally, we added 2,706 genome-wide significant GWAS SNPs from a recent blood trait GWAS[20]. SNP coordinates were lifted to hg19 using the *liftOver* command from R package rtracklayer v1.34.1[79] and subsequently standardized to match the GIANT 1kG p1v3 ALL reference panel. This yielded 10,562 SNPs (**Supplementary Table 2**). We tested associations between all risk factors and

genes that were at least 5Mb away to ensure that that they did not tag a *cis*-eQTL effect. All together, 10,317 trait-associated SNPs were tested in *trans*-eQTL analyses.

eQTS mapping

PGS trait inclusion

Full association summary statistics were downloaded from several publicly available resources (**Supplementary Table 13**). GWAS performed exclusively in non-European cohorts were omitted. Filters applied to the separate data sources are indicated in the **Supplementary Note**. All the dbSNP rs numbers were standardized to match GIANT 1kG p1v3, and the directions of effects were standardized to correspond to the GIANT 1kG p1v3 minor allele. SNPs with opposite alleles compared to GIANT alleles were flipped. SNPs with A/T and C/G alleles, tri-allelic SNPs, indels, SNPs with different alleles in GIANT 1kG p1v3 and SNPs with unknown alleles were removed from the analysis. Genomic control was applied to all the P-values for the datasets not genotyped by Immunochip or Metabochip. Additionally, genomic control was skipped for one dataset that did not have full associations available[80] and for all the datasets from the GIANT consortium, as for these genomic control had already been applied. All together, 1,263 summary statistic files were added to the analysis. Information about the summary statistics files can be found in the **Supplementary Note** and **Supplementary Table 14**.

PGS calculation

A custom Java program, GeneticRiskScoreCalculator-v0.1.0c, was used for calculating several PGS in parallel. Independent effect SNPs for each summary statistics file were identified by double-

clumping by first using a 250kb window and subsequently a 10Mb window with LD threshold R^2=0.1. Subsequently, weighted PGS were calculated by summing the risk alleles for each independent SNP, weighted by its GWAS effect size (beta or log(OR) from the GWAS study). Four GWAS P-value thresholds ($P<5\times10^{-8}$, 1×10^{-5}, 1×10^{-4} and 1×10^{-3}) were used for constructing PGS for each summary statistics file.

Pruning the SNPs and PGS

To identify a set of independent genetic risk factors, we conducted LD-based pruning as implemented in PLINK 1.9[81] with the setting --indep-pairwise 50 5 0.1. This yielded in 4,586 uncorrelated SNPs (R^2<0.1, GIANT 1kG p1v3 ALL).

To identify the set of uncorrelated PGS, ten permuted *trans*-eQTL Z-score matrices from the combined *trans*-eQTL analysis were first confined to the pruned set of SNPs. Those matrices were then used to identify 3,042 uncorrelated genes, based on Z-score correlations (absolute Pearson R < 0.05). Next, permuted eQTS Z-score matrices were confined to uncorrelated genes and used to calculate pairwise correlations between all genetic risk scores to define a set of 1,873 uncorrelated genetic risk scores (Pearson R^2 < 0.1).

Empirical probe matching

To integrate different expression platforms (four different Illumina array models, RNA-seq, Affymetrix U291 and Affymetrix Hu-Ex v1.0 ST) for the purpose of meta-analysis, we developed an empirical probe-matching approach. We used the pruned set of SNPs to conduct per-platform meta-analyses for all Illumina arrays, for all RNA-seq datasets, and for each Affymetrix dataset separately, using summary statistics from analyses without any gene expression correction for

principal components. For each platform, this yielded an empirical *trans*-eQTL Z-score matrix, as well as ten permuted Z-score matrices, where links between genotype and expression files were permuted. Those permuted Z-score matrices reflect the gene-gene or probe-probe correlation structure.

We used RNA-seq permuted Z-score matrices as a gold standard reference and calculated for each gene the Pearson correlation coefficients with all the other genes, yielding a correlation profile for each gene. We then repeated the same analysis for the Illumina meta-analysis, and the two different Affymetrix platforms. Finally, we correlated the correlation profiles from each array platform with the correlation profiles from RNA-seq. For each array platform, we selected the probe showing the highest Pearson correlation with the corresponding gene in the RNA-seq data and treated those as matching expression features in the combined meta-analyses. This yielded 19,960 genes that were detected in RNA-seq datasets and tested in the combined meta-analyses. Genes and probes were matched to Ensembl v71[82] (see **URLs**) stable gene IDs and HGNC symbols in all the analyses.

Cross-platform replications

To test the performance of the empirical probe-matching approach, we conducted discovery *cis*-, *trans*- and eQTS meta-analyses for each expression platform (RNA-seq, Illumina, Affymetrix U291 and Affymetrix Hu-Ex v1.0 ST arrays; array probes matched to 19,960 genes by empirical probe matching). For each discovery analysis, we conducted replication analyses in the three remaining platforms, observing strong replication of both *cis*-eQTLs, *trans*-eQTLs and eQTS in different platforms, with very good concordance in allelic direction.

Meta-analyses

We meta-analyzed the results using a weighted Z-score method[1], where the Z-scores are weighted by the square root of the sample size of the cohort. For *cis*-eQTL and *trans*-eQTL meta-analyses, this resulted in a final sample size of N=31,684. The combined eQTS meta-analysis included the subset of unrelated individuals from the Framingham Heart Study, resulting in a combined sample size of 28,158.

Quality control of the meta-analyses

For quality control of the overall meta-analysis results, MAFs for all tested SNPs were compared between eQTLGen and 1kG p1v3 EUR (**Extended Data Figure 3**), and the effect direction of each dataset was compared against the meta-analyzed effect (**Extended Data Figure 2A-C**).

FDR calculation for *trans*-eQTL and eQTS mapping

To determine nominal P-value thresholds corresponding to FDR=0.05, we used the pruned set of SNPs for *trans*-eQTL mapping and permutation-based FDR calculation, as described previously[1]. We leveraged those results to determine the P-value threshold corresponding to FDR=0.05 and used this as a significance level in *trans*-eQTL mapping in which all 10,317 genetic trait-associated SNPs were tested. In the eQTS analysis, an analogous FDR calculation was performed using a pruned set of PGSs. We analyzed only SNP/PGS-gene pairs tested in at least two cohorts.

Positive and negative set of *trans*-eQTLs

Based on the results of integrative *trans*-eQTL mapping, we defined true positive (TP) and true negative (TN) sets of *trans*-eQTLs. TP set was considered as all significant (FDR<0.05) *trans*-eQTLs. TN set of *trans*-eQTLs was selected as non-significant (max absolute meta-analysis Z-score 3; all FDR>0.05) SNP-gene combinations, adhering to following conditions:

1. The size of TN set was set equal to the size of TP set (59,786 *trans*-eQTLs).
2. Each SNP giving *trans*-eQTL effects on X genes in the TP set, is also giving *trans*-eQTL effects on X genes in the TN set.
3. Each gene that is affected in *trans* by Y SNPs in the TP set, is also affected in *trans* by Y SNPs in the TN set.
4. Adhere to the correlation structure of the SNPs: if two SNPs are in perfect LD, they affect the same set of genes, both in the TP set and in the TN set.
5. Adhere to the correlation structure of the genes: if two genes are perfectly co-expressed, they are affected by the same SNPs, both in the TP set and in the TN set.

This set of TN *trans*-eQTLs was used in subsequent enrichment analyses as the matching set for comparison.

Conditional *trans*-eQTL analyses

We aimed to estimate how many *trans*-eQTL SNPs were likely to drive both the *trans*-eQTL effect and the GWAS phenotype. The workflow of this analysis is shown in **Extended Data Figure 6**. We

used the integrative *trans*-eQTL analysis results as an input, confined ourselves to those effects which were present in the datasets we had direct access to (BBMRI-BIOS+EGCUT; N=4,339), and showed nominal $P < 8.3115\times 10^{-06}$ in the meta-analysis of those datasets. This P-value threshold was the same as in the full combined trans-eQTL meta-analysis and was based on the FDR=0.05 significance threshold identified from the analysis run on the pruned set of GWAS SNPs after removal of cross-mapping effects. We used the same methods and SNP filters as in the full combined *trans*-eQTL meta-analysis, aside from the FDR calculation, which was based on the full set of SNPs, instead of the pruned set of SNPs.

For each significant *trans*-eQTL SNP, we defined the locus by adding a ±1Mb window around it. Next, for each *trans*-eQTL gene we ran iterative conditional *trans*-eQTL analysis using all loci for given *trans*-eQTL gene. We then evaluated the LD between all conditional top *trans*-eQTL SNPs and GWAS SNPs using a 1 Mb window and $R^2>0.8$ (1kG p1v3 EUR) as a threshold for LD overlap.

Trans-eQTL mediation analysis

To identify potential mediators of *trans*-eQTL effects we used a G x E interaction model:

$t = \beta_0 + \beta_1 \times s + \beta_2 \times m + \beta_3 \times s \times m$

Where t is the expression of the *trans*-eQTL gene, s is the *trans*-eQTL SNP, and m is the expression of a potential mediator gene within 100kb of the *trans*-eQTL SNP. On top of the gene expression normalization that we used for the rest of our analysis, we used a rank-based inverse normal transformation to enforce a normal distribution before fitting the linear model, identical to the normalization used by Zhernakova et al.[47] in their G × E interaction eQTL analyses. We fitted this model separately on each of the cohorts that are part of the BIOS consortium. We transformed the interaction P-values to Z-scores and used the weighted Z-score method[83] to perform a meta-

analysis on the in total 3,831 samples. The Benjamini & Hochberg procedure[84] was used to limit the FDR to 0.05. The plots in **Extended Data Figure 9** are created with the default normalization, the regression lines are the best-fitting lines between the mediator gene and the *trans* eQTL gene, stratified by genotype. We used a Fisher's exact test to calculate the enrichment of significant (FDR ≤ 0.05) interactions between our TP *trans*-eQTLs and the interactions identified in the TN *trans*-eQTL set.

TF and tissue enrichment analyses

We downloaded the curated sets of known TF targets and tissue-expressed genes from the Enrichr web site[85,86]. TF target gene sets included TF targets as assayed by ChIP-X experiments from ChEA[87] and ENCODE[88,89] projects, and tissue-expressed genes were based on the ARCHS4 database[90]. Those gene sets were used to conduct hypergeometric over-representation analyses as implemented into the R package ClusterProfiler[91].

SMR analyses

To gain further insight into genes that are important in the biology of the trait, we used the combined *cis*-eQTL results to perform SMR[14] for 16 large GWAS studies (**Supplementary Table 20**). We derived *cis*-eQTL beta and standard error of the beta (SE(beta)) from the Z-score and the MAF reported in 1kG v1p3 ALL, using the following formulae[14]

$$\text{beta} = z / (\sqrt{(2p(1-p)(n+z^2))})$$

$SE(beta) = 1 / (\sqrt{(2p(1-p)(n+z^2))})$

Where p is the MAF and n is the sample size.

The *cis*-eQTLs were converted to the dense BESD format. The 1kG p1v3 ALL reference panel was also used to calculate LD, and SMR analysis was run using the SMR software v0.706 without any P-value cut-offs on either GWAS or eQTL input.

DEPICT

We applied DEPICT v194[15] to the same 16 recent GWAS traits as above (**Supplementary Table 20**), using all variants that attain a genome-wide significant P-value threshold. Specifically, we looked at the gene prioritization and gene set enrichment analyses to compare the results with the output of other prioritization methods (SMR [14]).

Comparison of gene prioritization with DEPICT and SMR

To investigate the consistency between results from two gene prioritization methods, we compared the enrichment of overlapping genes for 16 GWAS traits (**Supplementary Table 20**). We confined ourselves to genes that were tested in SMR and that fell within the DEPICT loci, and tested whether genes significant in SMR (P-value < 0.05 / number of tested genes) and DEPICT (FDR < 0.05) were enriched (one-sided Fisher's Exact Test).

Epigenetic marks enrichment

We ascertained epigenetic properties of the lead *cis*-eQTL SNPs, and contrasted these to a set of 3,688 trait-associated SNPs that were associated with either blood-related traits (such as mean corpuscular volume or platelet counts) or immune-mediated diseases. The SNPs were annotated with histone and chromatin marks information from the Epigenomics Roadmap Project. We summarized the information by calculating the overlap ratio across 127 human cell types between the epigenetic marks and the SNP within a window size of +/- 25bp: if a SNP co-localizes with a mark for all 127 cell-types, the score for that SNP will be 1; if a SNP co-localizes with a mark for none of the cell-types, the score will be 0.

The reason we chose only SNPs associated to blood-related traits and immune-mediated diseases was to minimize potential confounding due to a subtle bias in the Epigenomics Roadmap Project towards blood cell-types: 29 of the 127 cell-types that we studied were blood cell types. However, when redoing the epigenetic enrichment analysis, while excluding these blood cell types, we did not see substantial differences in the enriched and depleted histone marks.

Chromosomal contact analyses

Capture Hi-C overlap for *cis*-eQTLs

To assess whether *cis*-eQTL lead SNPs overlapped with chromosomal contact as measured using Hi-C data, we used promoter capture Hi-C data[92], downloaded from CHiCP[93] (see **URLs**). We took the lead eQTL SNPs and overlapped these with the capture Hi-C data and studied the 10,428 cis-eQTL genes for which this data is available. We then checked whether the Capture Hi-C target

maps within 5kb of the lead SNP. Of 508 *cis*-eQTL genes that mapped over 100 kb from the TSS or TES, 223 overlapped capture Hi-C data (27.8%). Of 7,984 *cis*-eQTL genes that mapped within 100kb from the TSS or TES, 1,641 overlapped capture Hi-C data (17.0%, Chi^2 test $P = 10^{-14}$). To ensure this was not an artefact, we performed the same analysis, while flipping the location of the capture Hi-C target with respect to the location of the bait, and did not observe any significant difference (Chi^2 test $P = 0.59$).

Hi-C overlap enrichment analysis for *trans*-eQTLs

To assess whether *trans*-eQTLs were enriched for chromosomal contacts as measured using Hi-C data, we downloaded the contact matrices for the human lymphoblastoid GM12878 cell line[19] (GEO accession GSE63525). We used the intrachromosomal data at a resolution of 10kb with mapping quality of 30 or more (MAPQGE30), and normalized using the KRnorm vectors. For each of the 59,786 *trans*-eQTLs, we evaluated whether any contact was reported in this dataset. We divided each *trans*-eQTL SNP and any of their proxies (R^2>0.8, 1kG p1v3, EUR, acquired from SNiPA[94]; **URLs**) in 10kb blocks. The *trans*-eQTL genes were also assigned to 10kb blocks, and to multiple blocks if the gene was more than 10kb in length (length between TSS and TES, Ensembl v71). For each individual *trans*-eQTL SNP-gene pair, we then determined if there was any overlap with the Hi-C contact matrices. We repeated this analysis using the true negative set of *trans*-eQTLs described before to generate a background distribution of expected contact.

Data availability

Full summary statistics from eQTLGen meta-analyses are available on the eQTLGen website: www.eqtlgen.org which was built using the MOLGENIS framework[95].

Code availability

Individual cohorts participating in the study followed the analysis plans as specified in the **URL**s or with slight alterations as described in the **Methods** and **Supplementary Note**. All tools and source code, used for genotype harmonization, identification of sample mixups, eQTL mapping, meta-analyses and for calculating polygenic scores are freely available at https://github.com/molgenis/systemsgenetics/.

Acknowledgments

The cohorts participating in this study list the acknowledgments in the cohort-specific supplemental information in Supplementary Note.

We thank i2QTL CONSORTIUM for providing the iPSC replication results.

We thank Kate McIntyre for editing the final text.

This work is supported by a grant from the European Research Council (ERC Starting Grant agreement number 637640 ImmRisk) to LF and a VIDI grant (917.14.374) from the Netherlands Organisation for Scientific Research (NWO) to LF. We thank the UMCG Genomics Coordination Center, MOLGENIS team, the UG Center for Information Technology, and the UMCG research IT program and their sponsors in particular BBMRI-NL for data storage, high performance compute and web hosting infrastructure. BBMRI-NL is a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO) [grant number 184.033.111].

URLs

Full summary statistics from this study, www.eqtlgen.org

ExAC pLI scores, http://exac.broadinstitute.org/downloads;

Ensembl v71 annotation file,

ftp://ftp.ensembl.org/pub/release-71/gtf/homo_sapiens;

Reference for genotype harmonizing,

ftp://share.sph.umich.edu/1000genomes/fullProject/2012.03.14/GIANT.phase1_release_v3.2010

1123.snps_indels_svs.genotypes.refpanel.ALL.vcf.gz.tgz

eQTLGen analysis plan for Illumina array datasets,

https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook;

eQTLGen analysis plan for RNA-seq datasets,

https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook-for-RNA-seq-data;

eQTLGen analysis plan for Affymetrix array datasets,

https://github.com/molgenis/systemsgenetics/wiki/QTL-mapping-analysis-cookbook-for-Affymetrix-expression-arrays;

GenotypeHarmonizer, https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer;

Protocol to resolve sample mixups, https://github.com/molgenis/systemsgenetics/wiki/Resolving-mixups;

Enrichr gene set enrichment libraries,

http://amp.pharm.mssm.edu/Enrichr/;

GeneOverlap package for enrichment analyses,

https://www.bioconductor.org/packages/release/bioc/html/GeneOverlap.html;

SHRiMP aligner used for re-mapping Illumina probes,

http://compbio.cs.toronto.edu/shrimp/;

EBI GWAS Catalogue,

https://www.ebi.ac.uk/gwas/;

Immunobase,

http://www.immunobase.org/;

ClusterProfiler package used for tissue enrichment analyses,

http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html;

Capture Hi-C data,

https://www.chicp.org/

SNiPA, used to acquire proxy SNPs,

http://snipa.helmholtz-muenchen.de/snipa3/

Regulatory Circuits, used to acquire TF data,

www.RegulatoryCircuits.org

References:

1. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
2. Kirsten, H. *et al.* Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci†. *Hum. Mol. Genet.* **24,** 4746–4763 (2015).
3. Lloyd-Jones, L. R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am. J. Hum. Genet.* **100,** 228–237 (2017).
4. Jansen, R. *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum. Mol. Genet.* **26,** 1444–1451 (2017).
5. Joehanes, R. *et al.* Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18,** 16 (2017).
6. Brynedal, B. *et al.* Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am. J. Hum. Genet.* **100,** 581–591 (2017).
7. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *Am. J. Hum. Genet.* **100,** 571–580 (2017).
8. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
9. Lewis, C. M. & Vassos, E. Prospects for using risk scores in polygenic medicine. *Genome Med.* **9,** 96 (2017).
10. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention

Setting. *Circulation* **135,** 2091–2101 (2017).

11. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
12. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18,** 86 (2017).
13. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
14. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48,** 481–487 (2016).
15. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).
16. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10,** e1004383 (2014).
17. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. methods* **13,** 366–370 (2016).
18. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. methods* **14,** 61–64 (2017).
19. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
20. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).
21. Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514,** 92–97 (2014).

22. Ding, Y. *et al.* ZNF131 suppresses centrosome fragmentation in glioblastoma stem-like cells through regulation of HAUS5. *Oncotarget* **8,** 48545–48562 (2017).

23. Oh, Y. & Chung, K. C. Small ubiquitin-like modifier (SUMO) modification of zinc finger protein 131 potentiates its negative effect on estrogen signaling. *J. Biol. Chem.* **287,** 17517–17529 (2012).

24. Lemaitre, R. N. *et al.* Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet.* **7,** e1002193 (2011).

25. Deelen, P. *et al*. Improving the diagnostic yield of exome-sequencing, by predicting gene-phenotype associations using large-scale gene expression analysis. bioRxiv preprint (2018).

26. Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7,** e1002216 (2011).

27. 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).

28. Gateva, V. *et al.* A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1228–1233 (2009).

29. Yin, X. *et al.* Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. *Nat. Commun.* **6,** 6916 (2015).

30. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *New Engl. J. Med.* **363,** 1211–1221 (2010).

31. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-

eQTLs and co-expression QTLs. *Nat. Genet.* **50,** 493–497 (2018).

32. Wang, J. H. *et al.* Aiolos regulates B cell activation and maturation to effector state. *Immunity* **9,** 543–553 (1998).
33. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).
34. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
35. He, M. *et al.* Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* **24,** 1791–1800 (2015).
36. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267,** 1360 LP-1363 (1995).
37. Chong, J. A. *et al.* REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80,** 949–957 (1995).
38. Cheong, A. *et al.* Downregulated REST transcription factor is a switch enabling critical potassium channel expression and cell proliferation. *Mol. cell* **20,** 45–52 (2005).
39. Dibner, C., Schibler, U. & Albrecht, U. The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu. Rev. Physiol.* **72,** 517–549 (2010).
40. Bass, J. & Lazar, M. A. Circadian time signatures of fitness and disease. *Sci.* **354,** 994–999 (2016).
41. Song, C. *et al.* REV-ERB agonism suppresses osteoclastogenesis and prevents ovariectomy-induced bone loss partially via FABP4 upregulation. *FASEB J. : Off. Publ. Fed. Am. Soc. Exp. Biol.* **32,** 3215–3228 (2018).
42. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic

inheritance. *bioRxiv* (2018).

43. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
44. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).
45. Wang, X. *et al.* Macrophage ABCA1 and ABCG1, but not SR-BI, promote macrophage reverse cholesterol transport in vivo. *J. Clin. Investig.* **117,** 2216–2224 (2007).
46. Goldstein, J. L. & Brown, M. S. Binding and degradation of low density lipoproteins by cultured human fibroblasts. Comparison of cells from a normal subject and from a patient with homozygous familial hypercholesterolemia. *J. Biol. Chem.* **249,** 5153–5162 (1974).
47. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49,** 139–145 (2017).
48. Singh, A. B., Kan, C. F. K., Shende, V., Dong, B. & Liu, J. A novel posttranscriptional mechanism for dietary cholesterol-mediated suppression of liver LDL receptor expression. *J. Lipid Res.* **55,** 1397–1407 (2014).
49. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42,** 441–447 (2010).
50. Kõks, S. & Kõks, G. Activation of GPR15 and its involvement in the biological effects of smoking. *Exp. Biol. Med.* **242,** 1207–1212 (2017).
51. van Iterson, M., van Zwet, E. W., BIOS Consortium & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18,** 19 (2017).
52. Bauer, M., Fink, B., Seyfarth, H.-J., Wirtz, H. & Frille, A. Tobacco-smoking induced GPR15-

expressing T cells in blood do not indicate pulmonary damage. *BMC Pulm. Med.* **17,** 159 (2017).

53. Kõks, G. *et al.* Smoking-Induced Expression of the GPR15 Gene Indicates Its Potential Role in Chronic Inflammatory Pathologies. *Am. J. Pathol.* **185,** 2898–2906 (2015).
54. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
55. Smirnova, T., Miniou, P., Viegas-Pequignot, E. & Mallet, J. Assignment of the human syntaxin 1B gene (STX) to chromosome 16p11.2 by fluorescence in situ hybridization. *Genomics* **36,** 551–553 (1996).
56. Sousa, I. *et al.* Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Mol. Autism* **1,** 7 (2010).
57. Agarwal, A. *et al.* Dysregulated expression of neuregulin-1 by cortical pyramidal neurons disrupts synaptic plasticity. *Cell reports* **8,** 1130–1145 (2014).
58. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).
59. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7,** 11122 (2016).
60. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46,** 543–550 (2014).
61. El-Hattab, A. W. Serine biosynthesis and transport defects. *Mol. Genet. Metab.* **118,** 153–159 (2016).
62. Leuzzi, V., Alessandrì, M. G., Casarano, M., Battini, R. & Cioni, G. Arginine and glycine

stimulate creatine synthesis in creatine transporter 1-deficient lymphoblasts. *Anal. Biochem.* **375,** 153–155 (2008).

63. Hart, C. E. *et al.* Phosphoserine aminotransferase deficiency: a novel disorder of the serine biosynthesis pathway. *Am. J. Hum. Genet.* **80,** 931–937 (2007).
64. Klomp, L. W. *et al.* Molecular characterization of 3-phosphoglycerate dehydrogenase deficiency--a neurometabolic disorder associated with reduced L-serine biosynthesis. *Am. J. Hum. Genet.* **67,** 1389–1399 (2000).
65. Shaheen, R. *et al.* Neu-Laxova syndrome, an inborn error of serine metabolism, is caused by mutations in PHGDH. *Am. J. Hum. Genet.* **94,** 898–904 (2014).
66. McLaughlin, H. M. *et al.* A recurrent loss-of-function alanyl-tRNA synthetase (AARS) mutation in patients with Charcot-Marie-Tooth disease type 2N (CMT2N). *Hum. Mutat.* **33,** 244–253 (2012).
67. Auer-Grumbach, M. Hereditary sensory neuropathy type I. *Orphanet J. rare Dis.* **3,** 7 (2008).
68. Hanada, K. Serine palmitoyltransferase, a key enzyme of sphingolipid metabolism. *Biochim. et Biophys. Acta* **1632,** 16–30 (2003).
69. Glinton, K. E. *et al.* Disturbed phospholipid metabolism in serine biosynthesis defects revealed by metabolomic profiling. *Mol. Genet. Metab.* **123,** 309–316 (2018).
70. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proc. Natl. Acad. Sci. United States Am.* **100,** 2610–2615 (2003).
71. Bennett, L. *et al.* Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *J. Exp. Med.* **197,** 711–723 (2003).
72. Pantel, A. *et al.* Direct type I IFN but not MDA5/TLR3 activation of dendritic cells is required

for maturation and metabolic shift to glycolysis after poly IC stimulation. *PLoS Biol.* **12,** e1001759 (2014).

73. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99,** 1245–1260 (2016).
74. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. notes* **7,** 901 (2014).
75. Rumble, S. M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5,** e1000386 (2009).
76. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
77. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinforma.* **27,** 2104–2111 (2011).
78. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids Res.* **45,** D896–D901 (2017).
79. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25,** (2009).
80. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48,** 1031–1036 (2016).
81. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).
82. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic acids Res.* **46,** D754–D761 (2018).
83. Zaykin, D. V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J. Evol. Biol.* **24,** 1836–1841 (2011).

84. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

85. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14,** 128 (2013).

86. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids Res.* **44,** W90–W97 (2016).

87. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinforma.* **26,** 2438–2444 (2010).

88. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Sci.* **306,** 636–640 (2004).

89. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).

90. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9,** 1366 (2018).

91. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : J. Integr. Biol.* **16,** 284–287 (2012).

92. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

93. Schofield, E. C. *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinforma.* **32,** 2511–2513 (2016).

94. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmuller, G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31,** (2015).

95. Swertz, M. *et al.* The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinforma.* **11,** 12 (2010).

Davenport et al. Genome Biology (2018) 19:168
https://doi.org/10.1186/s13059-018-1560-8

Genome Biology

RESEARCH Open Access

Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial

Emma E. Davenport[1,2,3,4], Tiffany Amariuta[1,2,3,4,5], Maria Gutierrez-Arcelus[1,2,3,4], Kamil Slowikowski[1,2,3,4,5], Harm-Jan Westra[1,2,3,4], Yang Luo[1,2,3,4], Ciyue Shen[6], Deepak A. Rao[7], Ying Zhang[8], Stephen Pearson[9], David von Schack[8], Jean S. Beebe[8], Nan Bing[8], Sally John[10], Michael S. Vincent[8], Baohong Zhang[8] and Soumya Raychaudhuri[1,2,3,4,5,11,12*]

Abstract

Background: Cytokines are critical to human disease and are attractive therapeutic targets given their widespread influence on gene regulation and transcription. Defining the downstream regulatory mechanisms influenced by cytokines is central to defining drug and disease mechanisms. One promising strategy is to use interactions between expression quantitative trait loci (eQTLs) and cytokine levels to define target genes and mechanisms.

Results: In a clinical trial for anti-IL-6 in patients with systemic lupus erythematosus, we measure interferon (IFN) status, anti-IL-6 drug exposure, and whole blood genome-wide gene expression at three time points. We show that repeat transcriptomic measurements increases the number of *cis* eQTLs identified compared to using a single time point. We observe a statistically significant enrichment of in vivo eQTL interactions with IFN status and anti-IL-6 drug exposure and find many novel interactions that have not been previously described. Finally, we find transcription factor binding motifs interrupted by eQTL interaction SNPs, which point to key regulatory mediators of these environmental stimuli and therefore potential therapeutic targets for autoimmune diseases. In particular, genes with IFN interactions are enriched for ISRE binding site motifs, while those with anti-IL-6 interactions are enriched for IRF4 motifs.

Conclusions: This study highlights the potential to exploit clinical trial data to discover in vivo eQTL interactions with therapeutically relevant environmental variables.

Keywords: eQTL, Interactions, Clinical trials, Cytokines

Background

Cytokines are critical signals used by the immune system to coordinate inflammatory responses. These factors bind to specific receptors to induce widespread transcriptional effects. Cytokines and their receptors are not only genetically associated with susceptibility to a range of human diseases; they have also emerged as effective therapeutic targets [1]. Blockade of tumor necrosis factor (TNF) was the first cytokine-directed therapy to achieve widespread use and is now used broadly to treat multiple inflammatory diseases including rheumatoid arthritis (RA), psoriasis, and inflammatory bowel disease [2]. More recently, IL-6 has emerged as a compelling therapeutic target. IL-6 levels are elevated in autoimmune diseases such as systemic lupus erythematosus (SLE) and RA. The IL-6 receptor has been successfully targeted with tocilizumab in RA [3] and giant cell arteritis [4], while IL-6 has been targeted directly with siltuximab for successful treatment of Castleman's disease [5]. In SLE, IL-6 is thought to play a role in the observed B cell hyperactivity and autoantibody production [6]. Targeting IL-6-R in SLE has shown promise in phase I trials [7], and this has led to the development of other biologics targeting IL-6 such as PF-04236921 [8]. Interferon (IFN)-α, produced primarily by plasmacytoid dendritic cells, has pleiotropic effects on the immune system. It has been implicated as a key mechanism in SLE development and pathogenesis and is being investigated as a therapeutic target [9]. Agents targeting other inflammatory cytokines, including interleukin-1 (IL-1), IL-12,

* Correspondence: soumya@broadinstitute.org
[1]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA 02115, USA
[2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
Full list of author information is available at the end of the article

IL-17A, and IL-23 are also in clinical use to treat autoimmune conditions. Interestingly, IL-1 blockade with canakinumab has also been recently reported to reduce risk of heart attacks, stroke, and cardiovascular disease [10]. Therefore, defining the regulatory consequences of physiologic perturbations of cytokine levels will inform our understanding of both disease and drug mechanisms.

A *cis* expression quantitative trait locus (eQTL) contains a genetic variant that alters expression of a nearby gene. *Cis* eQTLs are ubiquitous across the genome [11], and while most are stable across tissues and conditions, environmental variables can alter the effects of some of them [12–18]. If an environmental change leads to disruption of regulators upstream of a gene, then it could magnify or dampen an eQTL effect, resulting in a genotype-by-environment interaction (Additional file 1: Figure S1). Therefore, observing a set of eQTL interactions due to a perturbagen, such as a cytokine, can identify shared upstream regulatory mechanisms, such as transcription factors and key pathways. Alternatively, a set of shared eQTL interactions may be the consequence of a cellular subpopulation whose frequency is being altered by the perturbagen. Even a single eQTL interaction where we can define mechanism can lead to insights about the action of the perturbagen.

However, *cis* eQTL interactions with physiologic environmental factors in humans have been challenging to discover in vivo [19–23] even with large cohorts [11, 17]. Success at finding *cis* eQTL interactions has largely been found in studies using model organisms [24, 25] or treating cells in vitro with non-physiologic conditions [26]. Thus far, these studies might be limited in power since they often map eQTLs separately across conditions and fail to exploit the power of repeat measurements [27]. In other instances, they test for genetic variants associated with differential expression and miss information about the magnitude of the eQTL effect in a specific condition [28].

We predicted that if the transcriptome is assayed at multiple time points under different exposure states, then the repeat measurements could lead to an increase in power to detect eQTLs and their interactions with environmental perturbations. If the same individual is assessed at multiple times, then the noise in transcriptomic measurements is reduced. Furthermore, repeat measurements from the same individuals when they are both unexposed and exposed to an environmental perturbagen allow for more accurate modeling of the effect of the perturbagen within those subjects.

Clinical trials, with their structured study design, may be the ideal setting to detect eQTL interactions with therapeutically important variables. In clinical trials, it is becoming increasingly common to collect transcriptional and genetic data alongside clinical and physiological data [29]. This extensive phenotyping of therapeutically important variables and biomarkers within the same individual at multiple time points provides a unique opportunity to identify in vivo eQTL interactions.

Here, we examined the modulation of eQTL effects by environmental factors that alter cytokine levels using data from a phase II clinical trial to evaluate the safety and efficacy of a neutralizing IL-6 monoclonal antibody (PF-04236921) in 157 SLE patients [8] ("Study design"). Many patients with SLE exhibit high levels of genes induced by type I IFN; these genes, known as the IFN signature, are a marker of disease severity [30, 31] and a pathogenic feature of SLE. This feature of the disease, together with exposure to anti-IL-6 leads to cytokine fluctuations in this cohort yielding opportunities to assess the impact of cytokine levels on eQTL effects. While this drug was not significantly different from placebo for the primary efficacy endpoint (proportion of patients achieving the SLE Responder Index (SRI-4) at week 24), biologically it effectively reduced free IL-6 protein levels (Additional file 1: Figure S2). Given the key role of IL-6 and IFN in a range of diseases, the downstream regulatory effects of these cytokines are of great interest to study.

In this study, we leverage the power of repeat transcriptional and environmental measurements from a lupus clinical trial to identify in vivo eQTL interactions with IFN status and anti-IL-6 exposure. In the process, we define novel eQTL interactions for both IFN and IL-6.

Results

We conducted whole blood high-depth RNA-seq profiling at 0, 12, and 24 weeks in anti-IL-6 exposed and unexposed individuals with the Illumina TruSeq protocol. We quantified 20,253 gene features and examined 1,595,793 genotyped and imputed common variants genome-wide ("RNA sequencing", "Genotyping", "Imputation"). Along with each RNA-seq assay, we documented anti-IL-6 exposure and quantified IFN signature status with real-time PCR.

Mapping eQTL in SLE patients

We first mapped *cis* eQTLs and then tested them for interactions with IFN status and anti-IL-6 exposure. eQTL interactions can be explored using our interactive visualization tool (http://baohongz.github.io/Lupus_eQTL [32], Additional file 1: Figure S3).

To identify *cis* eQTLs, we examined the association between gene expression and SNPs within 250 kb upstream of the transcription start site and 250 kb downstream of the transcription end site. In order to account for repeat measurements, with up to three RNA-seq assays per patient (Fig. 1a, 379 samples from 157 patients, "eQTL and interaction analysis"), we used a linear mixed model. We

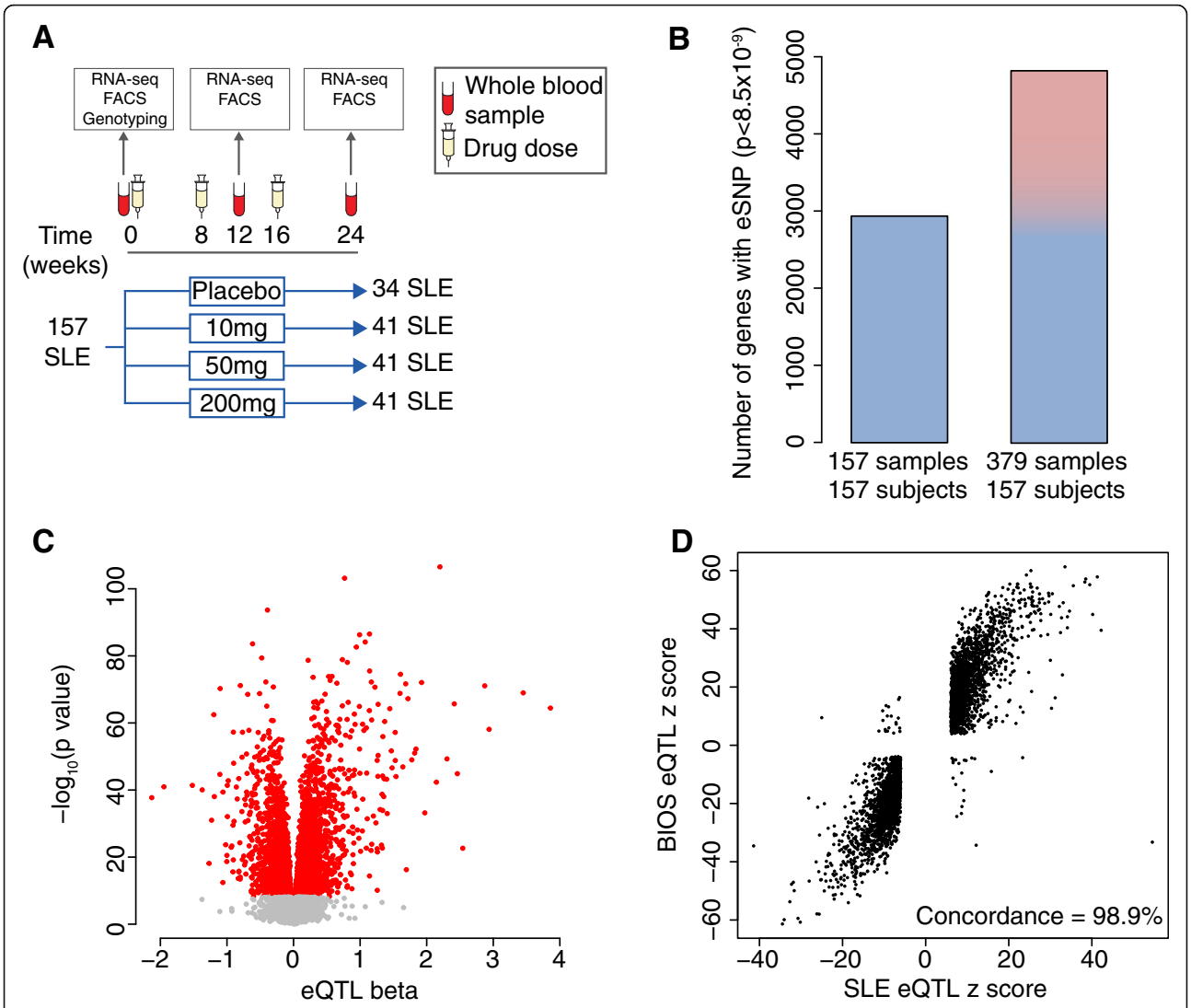


Fig. 1 Identifying eQTLs in SLE patients. **a** Clinical trial structure and sampling strategy for the individuals used for eQTL analysis. The samples available are summarized in Table 1. **b** Number of eQTL genes identified using a linear model (left) and a linear mixed model (right). For the linear model, we used the first available time point for each individual (week 0 sample for $n = 152$, week 12 sample for $n = 5$). **c** Volcano plot of eQTL effects for the most significantly associated SNP for each gene (red color indicates $p < 8.5 \times 10^{-9}$). **d** Concordance of SLE eQTL effects ($p < 8.5 \times 10^{-9}$) with eQTLs observed in the BIOS cohort [11] of healthy individuals (FDR < 0.05). Each point represents the most significant SNP-gene pair for the SLE eQTL

included 25 gene expression principal components to maximize the number of eQTL detected and 5 genotyping principal components to account for the heterogeneity in ethnicity in our cohort ("eQTL and interaction analysis"). We observed that the multi-ethnic nature of our study did not confound our results, consistent with Stranger et al. [33] (Additional file 1: Figure S4).

To ensure we only tested for interactions in a set of highly confident eQTLs, we applied a stringent correction for the total number of hypotheses tested. We recognized that this approach might arguably be overly stringent for eQTL discovery, but we wanted to be certain that we were only testing eQTLs for interactions that had a convincing main effect. Since we tested a total of 5,872,001 SNP-gene pairs genomewide, we set a significance threshold of $p_{eqtl} < 8.5 \times 10^{-9}$ (0.05/5,872,001 tests). We identified 4818 *cis* eQTL genes (Fig. 1b, c, Additional file 2: Table S1). The summary statistics for all the gene SNP pairs tested are available through figshare [34].

To confirm the validity of our eQTLs, we compared them to a larger dataset. In the BIOS cohort,

consisting of 2166 healthy individuals [11], we observed that 85.4% of our SLE eQTL SNP-gene pairs are reported as eQTLs (FDR < 0.05). Of these, 98.9% showed consistent direction of effect ($p < 5 \times 10^{-16}$, binomial test, Fig. 1d), suggesting that our results were highly concordant with those in this substantially larger study.

Repeat measurements increase power to detect eQTL

Under reasonable assumptions, we would expect repeat samples to increase our power. Supporting that expectation, we detected 64% more *cis* eQTLs compared to the 2934 genes from using a single sample (first available time point) per individual (Fig. 1b). An alternative might have been to identify eQTLs separately from each of the three time points; however, this approach identified only a total of 3050 eQTL genes (Additional file 1: Figure S5). Modeling all three time points together results in 58% more *cis* eQTLs than modeling each time point separately.

We speculated that while repeat measures did increase power over single measures, that given a fixed number of samples, independent samples would lead to more power. To this end, we conducted an analysis fixing the number of samples at 157 and using 53 individuals with repeat measures (with two missing samples). Unsurprisingly, we found fewer eQTLs (2215 genes) with the repeat measures alone compared to an analysis with the same number of independent samples (2934 genes).

IFN status eQTL interactions

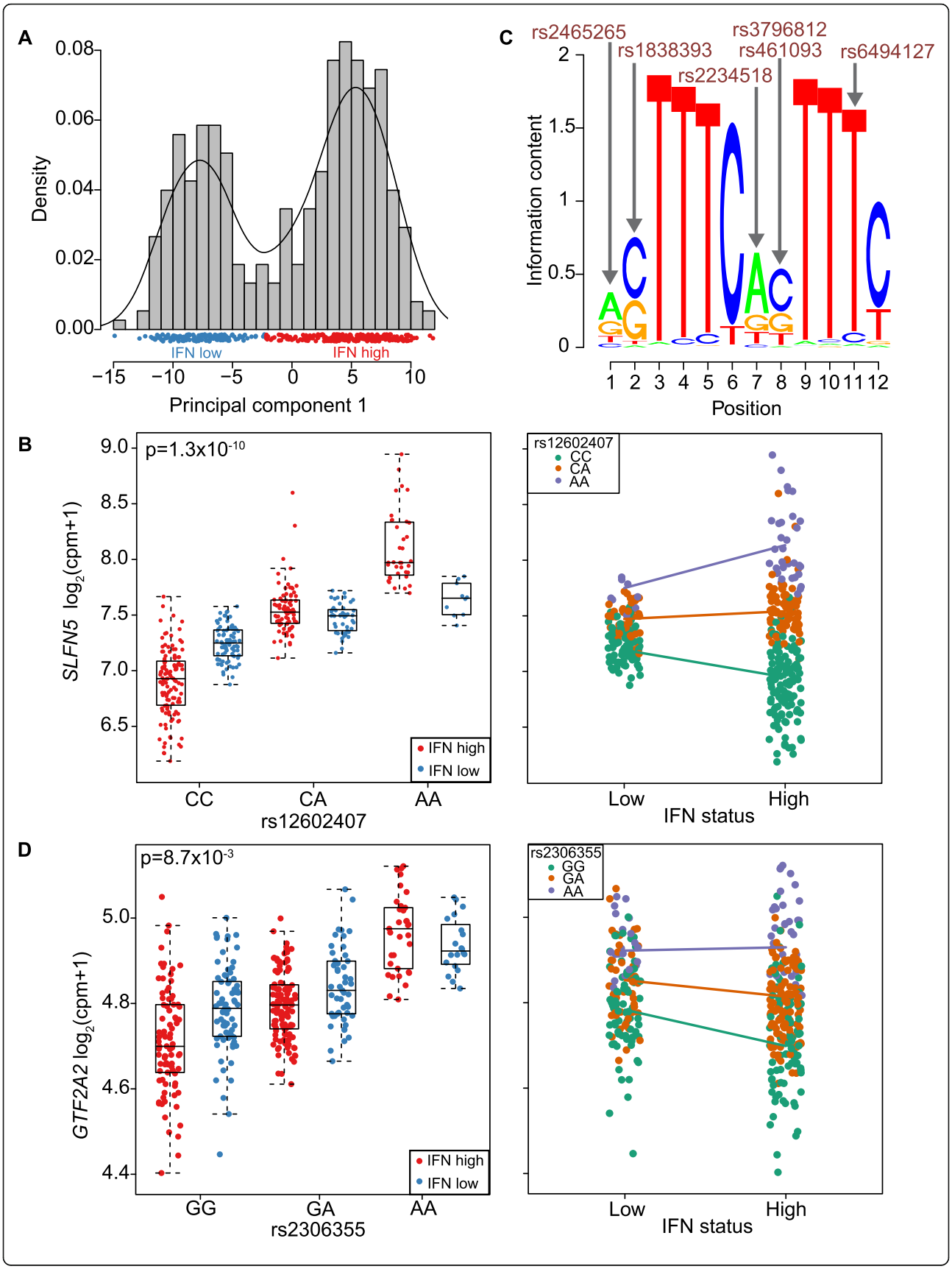
For each of the 4818 *cis* eQTL genes, we tested the most significantly associated SNP for environmental interactions with our linear mixed model framework. We first explored the influence of type I IFN on gene regulation after determining the IFN status of every patient at each time point. We classified each sample as either IFN high or IFN low using real-time PCR of 11 IFN-inducible genes [35] ("Interferon status", Fig. 2a).

We first wanted to assess whether our results were indeed enriched for interactions. To do this, we identified those eQTLs with nominally significant interaction effects at $p_{\text{interact}} < 0.01$. We would expect ~48 out of 4818 from chance alone. Surprisingly, we observed 182 IFN-eQTL interactions (Additional file 2: Table S1) that were nominally significant at $p_{\text{interact}} < 0.01$ suggesting that there was evidence of enrichment for eQTL interactions. We conducted permutations to ensure that these results were not the consequence of potentially inflated statistics, which might be the result for example of low-frequency alleles, genes violating normality assumptions, or other technical artifacts. In each of 1000 stringent permutations, we simply reassigned IFN status across samples and retested for eQTL interactions. This permutation preserves the main eQTL effect, since it maintains genotypes of the individuals with the associated expression data, but disrupts any real interactions that might be present in the data. In 0 out of 1000 instance did we observe 182 or more interactions at $p_{\text{interact}} < 0.01$ suggesting that the number of observed interactions is enriched and highly unlikely to have happened by chance (Additional file 1: Figure S6, $p_{\text{permute}} \sim 0/1000 = < 0.001$).

We then went on to identify those specific IFN-eQTL interactions of greatest interest by calculating a false discovery rate or *q* value for each interaction using the *q* value package [36] ("eQTL and interaction analysis"). We observed a total of 210 interactions with an FDR < 0.2 threshold (111 with FDR < 0.1 and 67 with FDR < 0.05, Additional file 2**:** Table S1). We note that 11 of these genes have already been described as having an interaction with a proxy gene for type I IFN signaling in the much larger BIOS study [11]. For example, *SLFN5* expression is influenced by the rs12602407 SNP ($p_{\text{interact}} = 1.3 \times 10^{-10}$, FDR < 9.9×10^{-8}, Fig 2b), and this effect is magnified in IFN high samples. Of these 210 IFN-eQTL interactions, 99 were not reported in the BIOS study [11]. Indeed, applying a more stringent cut off of FDR < 0.01, 27/34 of our interactions are not previously reported and therefore are almost certainly novel IFN-eQTL interactions with high confidence (Additional file 1: Figure S7).

We speculated that groups of eQTL interactions might be driven by the same common regulatory factor. We divided interactions into magnifiers, where the environmental exposure increases the size of the eQTL effect, and dampeners where the environmental exposure decreases the eQTL effect (Additional file 1: Figure S8). We hypothesized that the transcription factors driving the response to type I IFN may be different for the eQTL interactions defined as magnifiers ($n = 127$, FDR < 0.2) and dampeners ($n = 83$, FDR < 0.2).

We applied HOMER [37] to assess overlap between transcription factor binding motifs and the eQTL interaction SNPs (and SNPs in high linkage disequilibrium (LD, $r^2 > 0.8$) in the *cis* window, "eQTL and interaction analysis"). To determine enrichment, we compared the transcription factor motifs found in a set of sequences (containing the interaction SNPs) from one category of interactions relative to the other. We conducted two separate analyses: the proportion of magnifying eQTL interaction sequences with a motif compared to the proportion of dampening interaction sequences with a motif and vice versa. We found enrichment of motifs for key transcription factors involved in IFN signaling including a statistically significant enrichment for the ISRE motif (HOMER $p = 1 \times 10^{-4}$, Additional file 2: Table S2). The ISRE motif disruption occurred for 11 genes with an


A
Density
0.08
0.06
0.04
0.02
0.00
IFN low
IFN high
−15 −10 −5 0 5 10
Principal component 1
C
rs2465265
rs1838393
rs2234518
rs3796812
rs461093
rs6494127
Information content
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12
Position
B
p=1.3x10-10
SLFN5 log2(cpm+1)
9.0
8.5
8.0
7.5
7.0
6.5
IFN high
IFN low
CC CA AA
rs12602407
rs12602407
CC
CA
AA
Low High
IFN status
D
p=8.7x10-3
GTF2A2 log2(cpm+1)
5.0
4.8
4.6
4.4
IFN high
IFN low
GG GA AA
rs2306355
rs2306355
GG
GA
AA
Low High
IFN status

(See figure on previous page.)**Fig. 2** eQTL interactions with IFN status. **a** Designation of IFN status for each sample from the real-time PCR expression of 11 genes (first principal component). **b** IFN status interaction with the *SLFN5* eQTL plotted with respect to rs12602407 genotype (left) and IFN status of the sample (right). **c** The ISRE motif enriched among eQTLs magnified in IFN high samples. Arrows indicate positions of the motif interrupted by interaction SNPs (or SNPs in strong LD). Red indicates these SNPs correspond to magnified eQTLs. **d** IFN status interaction with the *GTF2A2* eQTL plotted with respect to rs2306355 genotype (left) and IFN status of the sample (right)

eQTL magnified in IFN high samples but for only one gene with an eQTL dampened (permutation $p < 0.019$, "Magnifiers and dampeners", Fig. 2c). An example is the *GTF2A2* rs2306355 eQTL ($p_{\text{interact}} = 8.7 \times 10^{-3}$, FDR < 0.15, Fig. 2d); rs2306355 is in tight LD ($r^2 = 0.83$ in Europeans) with rs6494127, which interrupts the TTCNNT TT core of the ISRE motif (Fig. 2c). This SNP likely disrupts IRF9 and STAT2 binding in the ISGF3 complex [38], which binds to the ISRE motif. We observe greater expression of *GTF2A2* in individuals with the rs2306355 A allele compared to G; this difference is magnified in IFN high individuals (Fig. 2d).

We included principal components as covariates in our model to account for confounding sources of gene expression variation that are not limited to those that have been measured in the study. Additional file 2: Table S3 summarizes the correlation between the principal components and potential known confounders such as age, sex, and site of recruitment. As no single principal component strongly correlates with these known confounders, we re-ran the interaction analysis including age and sex as fixed effects and site as a random effect. The interaction betas are very highly correlated ($r_s = 0.99$) with the original effects suggesting the principal components are capturing these known confounders (Additional file 1: Figure S9).

We considered that the principal components included as covariates in our model might be mitigating power. For example, the 4th principal component of gene expression is correlated with the IFN signature status of the sample ($r_s = -0.7$, Additional file 2: Table S3), so we repeated the IFN interaction analysis without correcting for principal component 4. For all the eQTLs tested for an IFN interaction, we observed very similar results with highly correlated z-scores ($r_s = 0.94$, Additional file 1: Figure S10). To further explore this, we also repeated the IFN interaction analysis without correcting for any expression principal components. While we find the betas for the interaction term are highly correlated ($r_s = 0.88$, Additional file 1: Figure S11a), only 23/210 of our IFN eQTL interactions remain significant with an FDR < 0.2 without correcting for any expression principal components. This reduction in significant interactions is likely due to the larger standard errors of the interaction estimate that are observed when principal components are not corrected for (Additional file 1: Figure S11b). Furthermore, 107/210 of these interactions no longer have a main eQTL effect (passing our Bonferroni corrected p value threshold) without principal component correction, further reducing our power to detect significant interactions.

Discovery of eQTL interactions with anti-IL-6 drug exposure

We then examined whether IL-6 blockade alters the relationship between genomic variation and gene expression and induces drug-eQTL interactions. We wanted to first test if there was evidence of such interactions in our data set. Again, using a threshold of $p_{\text{interact}} < 0.01$ for nominal significance for interactions, we observed 121 drug-eQTL interactions with anti-IL-6 out of 4818 eQTLs tested (Additional file 2: Table S1); similar to IFN interactions, this is far in excess of the ~ 48 we would expect by chance. As above, to ensure that these results were not the consequence of statistical artifact, we applied the same stringent permutation strategy, reassigning which samples were exposed or not to anti-IL-6. After 1000 permutations, we never observed as many as 121 drug-eQTL interactions with $p_{\text{interact}} < 0.01$ (Additional file 1: Figure S12), suggesting that our eQTLs were indeed highly enriched for those interacting with anti-IL-6 ($p_{\text{permute}} \sim 0/1000 < 0.001$).

We analyzed drug and IFN-eQTL interactions independently because anti-IL-6 exposure and IFN status are not associated (Fisher's exact test $p = 0.6$). However, to further ensure that these variables are independent, we repeated the interaction analysis with a full model including the drug, drug interaction, IFN, and IFN interaction terms. We find that the interaction betas are highly concordant ($r_s = 0.99$) with the original analysis (Additional file 1: Figure S13) for both IFN and drug-eQTL interactions providing further evidence that IFN status and drug exposure are independent.

To identify specific eQTL events that interact with anti-IL-6, we again calculated a false discovery rate. We observed that 72 of these interactions have an FDR < 0.2 (7 with FDR < 0.1 and 1 with FDR < 0.05, Additional file 2: Table S1). Only eight of these drug-eQTL interactions overlap with the interactions observed for IFN status (Additional file 2: Table S1). We note biologically relevant drug-eQTL interactions for *IL10* ($p_{\text{interact}} = 2.6 \times 10^{-3}$, FDR < 0.19, Additional file 1: Figure S14), an anti-inflammatory cytokine, *CLEC4C* ($p_{\text{interact}} = 2.9 \times 10^{-3}$, FDR < 0.19) which has previously been associated in *trans* with an SLE risk allele [39] and *CLEC18A* ($p_{\text{interact}} = 5.1 \times 10^{-4}$, FDR < 0.14, Fig. 3a) another member of the C-type lectin domain family.

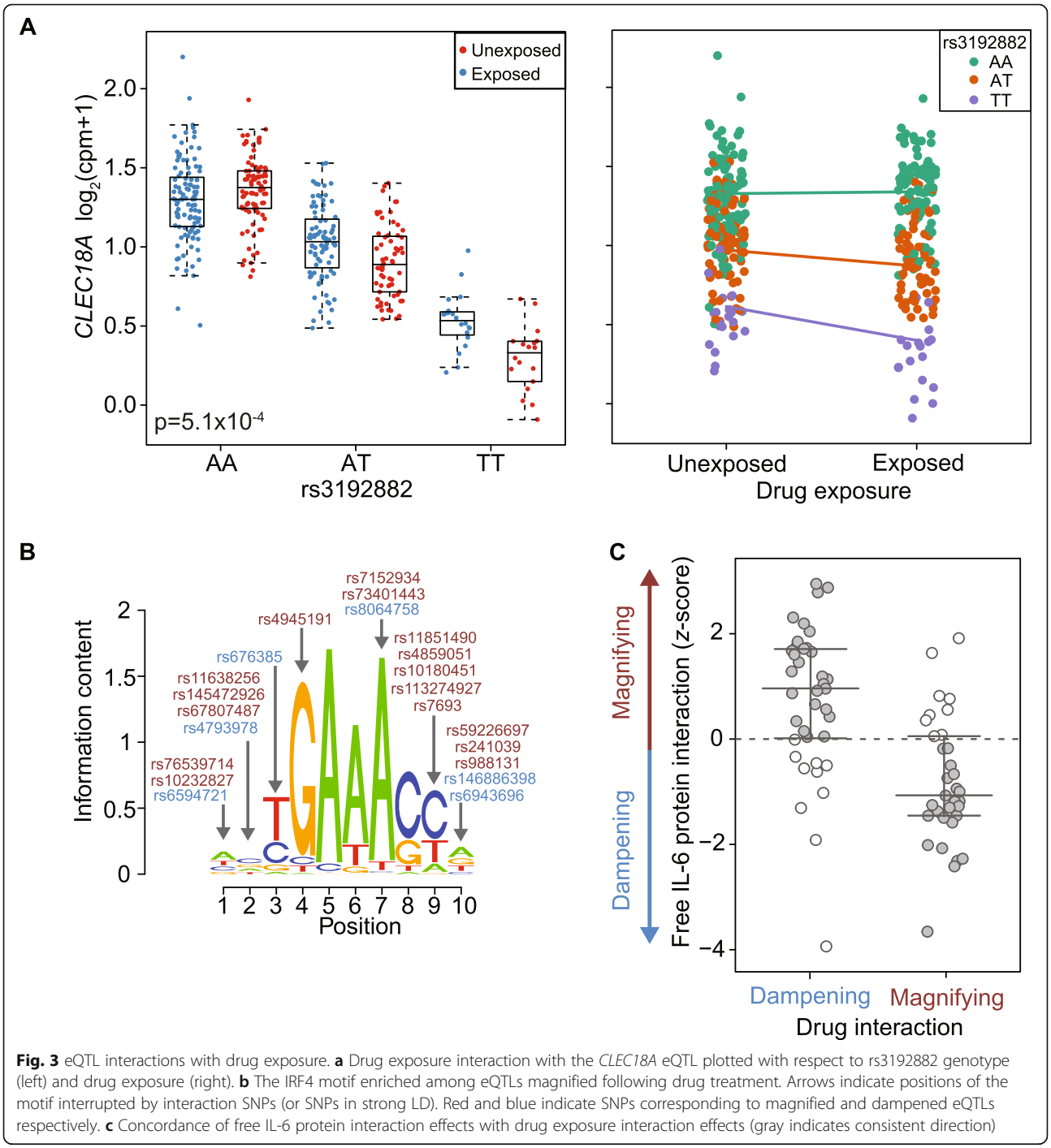


Fig. 3 eQTL interactions with drug exposure. **a** Drug exposure interaction with the *CLEC18A* eQTL plotted with respect to rs3192882 genotype (left) and drug exposure (right). **b** The IRF4 motif enriched among eQTLs magnified following drug treatment. Arrows indicate positions of the motif interrupted by interaction SNPs (or SNPs in strong LD). Red and blue indicate SNPs corresponding to magnified and dampened eQTLs respectively. **c** Concordance of free IL-6 protein interaction effects with drug exposure interaction effects (gray indicates consistent direction)

Similar to the IFN-eQTL interactions, we divided the drug-eQTL interactions into magnifiers ($n = 33$, FDR < 0.2) and dampeners ($n = 39$, FDR < 0.2) (Additional file 1: Figure S15) and used the approach as described above to define transcription factors potentially driving the response to IL-6 blockade (Additional file 2: Table S4). One of the motifs enriched for eQTLs magnified after drug treatment (compared to dampeners) was IRF4 (HOMER $p = 1 \times 10^{-3}$). The IRF4 motif disruption occurred for nine genes, including *CLEC18A*, with an eQTL magnified after drug treatment compared to four genes with an eQTL dampened (Fig. 3b, "Magnifiers and dampeners"). We permuted the magnifying and dampening genes and found this ratio for enrichment is interesting at the gene level but not significant ($p = 0.058$) and therefore additional eQTL interactions will be necessary to confirm.

Comparing differential expression to eQTL interactions

A more common strategy to determine the effect of an environmental variable is to use differential gene expression. For IFN status, we identified 1850 differentially expressed genes (FDR < 0.05 and fold change > 1.2, Additional file 1: Figure S16, Additional file 2: Table S5). Only 42/210 IFN-eQTL interaction genes also show evidence of differential expression. For differential expression following anti-IL-6 treatment, we identified 394 genes (FDR < 0.05 and fold change > 1.2, Additional file 1: Figure S16, Additional file 2: Table S6). Only 2/72 drug-eQTL interaction genes show evidence of differential gene expression. This suggests that eQTL interactions offer independent information from differential expression, which might contribute to defining mechanisms.

Concordance of drug-eQTL interactions with protein level interactions

We hypothesized that interactions due to drug exposure are likely driven by free IL-6 cytokine levels (our key clinical biomarker of interest). If this is the case, for eQTLs dampened by drug exposure, an increase in free IL-6 should elicit an opposite interaction effect and result in eQTL magnification. We assessed whether eQTL interactions with free IL-6 protein levels measured in the patient serum samples were consistent with those following IL-6 blockade. We observed enrichment in the overlap between cytokine interactions and drug interactions (53/72 interactions in expected opposite direction, Fig. 3c, $p = 3.8 \times 10^{-5}$, binomial test).

Contribution of cell proportions to eQTL interactions

Given that we have conducted our study on whole blood, the observed eQTL interactions could be the consequence of a cellular subpopulation whose frequency is being altered by the environmental perturbagen or variability in cell type proportions between individuals. To explore this, we first determined B and T cell abundance from FACS data ("Cell counts"). We find that IFN and exposure to anti-IL-6 are not correlated with either B or T cell abundance (Additional file 1: Figure S17) suggesting that the proportions of these particular cell types are not being altered by the shifts in cytokine levels.

We then went on to explore the effect of correcting for cell proportions on our eQTL interaction effects. As our FACS data do not cover all relevant cell types, we also inferred the relative proportions of nine hematopoietic populations from the RNA-seq data using CIBERSORT [40]. For IFN, the interaction betas remain highly correlated ($r_s = 0.99$, $r_s = 0.998$, Additional file 1: Figure S18) after correcting for either B and T cell proportions from the FACS data or the nine hematopoietic proportions from CIBERSORT. Furthermore, the majority of the interactions (162/210 and 189/210) remain significant (FDR < 0.2) after these respective corrections. We observe a similar pattern for the drug-eQTL interactions (Additional file 1: Figure S19) where the interaction betas are again highly correlated ($r_s = 0.98$ and $r_s = 0.993$ for FACS and CIBERSORT proportion correction respectively). However only 15/72 and 54/72 drug-eQTL interactions remain significant after these corrections. This reduction in significant interactions suggests that some of these interactions may be related to changes in cell proportions. However, given that the interaction betas are so highly correlated and the relatively small effect size of the drug interactions, this could also be the result of reduced power to detect interactions following the inclusion of additional covariates in the model.

Discussion

In this study we mapped eQTLs in a clinical trial of SLE patients and discovered interactions with IFN and IL-6, two clinically important cytokines. Our study had dramatic variation in IL-6 that was therapeutically induced, and variation in IFN due to the disease status of the SLE patients. This, together with the structured study design with repeat measurements of gene expression across different conditions in the same individual, allowed us to identify in vivo eQTL interactions.

eQTL interactions with drug interventions or other therapeutically relevant physiologic variables are important to identify as they can point to regulatory mechanisms, such as transcription factors or subclasses of enhancers, acting downstream of the environmental condition of interest and driving groups of eQTL interactions. The IFN status eQTL interactions we identified provide support for this approach. By making use of the direction of effect for the eQTL interaction, we were able to identify an enrichment of magnifying eQTL interaction SNPs interrupting the binding sites of transcription factors known to be important in the response to IFN, such as ISGF3 (the STAT1, STAT2, and IRF9 complex), which binds ISRE. Once we are able to recognize the downstream drivers of therapeutically relevant clinical variables, then it may become possible to define more mechanisms of action for drugs and more precise drug targets.

As a powerful example, we note enrichment of magnifying anti-IL-6-eQTL interaction SNPs interrupting the binding site of IRF4. It has been suggested that IRF4 works downstream of IL-6 by binding BATF and coordinately regulating the production of IL10 and other genes [41]. Consistent with this, we observed that the *IL10* eQTL does indeed interact with presence of anti-IL-6 (Additional file 1: Figure S14). Previous studies have highlighted a role for IRF4 in the pathogenesis of autoimmune diseases in mouse and humans. For example in

a murine model of SLE, *IRF4* knockout mice did not develop lupus nephritis [42]. In humans, IRF4 is associated with RA [43], a disease in which anti-IL-6 treatment has been successful [3]. Our findings provide further support that IRF4 could be a potential therapeutic target for autoimmune diseases such as RA where anti-IL-6 is effective [44].

The ability to focus on interactions with specific patient phenotypes might point to key targets for disease intervention. For example, IFN is a key immunophenotype in SLE patients, and elevated in SLE compared to healthy controls [30, 31]. The IFN status immunophenotype is already itself driving interest in therapeutic targets. A recent phase II clinical trial has shown that an antagonist to the type I IFN receptor, acting upstream of ISRE, reduced severity of symptoms in SLE. Interestingly, the antagonist was more effective in the patients with a high baseline IFN status [45]. This example provides a compelling case study for how understanding master regulators of key disease phenotypes might lead to promising new therapeutic strategies. We speculate that this provides a mechanism for stratified medicine for future studies, which may be applicable to other diseases.

We recognized that computing eQTL interactions requires a robust statistical model that accounts for genotype, environmental factor, RNA expression levels, repeat measurements, and technical covariates. We were sensitive to the possibility that pre-processing and normalization of these factors could potentially have an impact on our results. For this reason, we used stringent filtering and examined only variants that were common and where the minor allele was present for each of the exposure groups. Next, to confirm enrichment of eQTL interactions, we used a stringent permutation-based strategy that preserved the distribution of genotypes and corresponding expression values. Finally, we also utilized a standard normal transformation [46] ("eQTL and interaction analysis") and observed that this had little effect on the primary eQTL analysis ($r_s = 0.99$ for z scores, Additional file 1: Figure S20) and interaction analyses (IFN $r_s = 0.84$, drug $r_s = 0.76$ for z scores, Additional file 1: Figure S21), or the observed enrichment over the null in our stringent permutation analysis (Additional file 1: Figure S22).

We acknowledge that our approach for eQTL discovery using a stringent Bonferroni corrected p value threshold is conservative and could reduce our ability to detect eQTLs with a modest effect in one group and therefore reduce the number of interactions we observe. However, given the challenge of identifying interactions, we wanted to ensure that we were confident in the eQTL effect before testing that effect for an interaction. Furthermore, as demonstrated by the *SLFN5* IFN-eQTL interaction, we still observe interaction examples where an eQTL effect is very modest in one group, in this case, samples designated as IFN low (Fig. 2b).

While we find that the majority of the eQTL interactions that we identify are independent of differentially expressed genes, we have used the common strategy for identifying differential expression, which does not take into account the genotype of the individuals. The differential expression presented here therefore represents the average change in expression across all genotypes, regardless of any eQTL or interaction effect. Furthermore, like most differential expression approaches, we have employed a fold change cut-off. Using statistical evidence alone, 121/210 IFN-eQTL interaction genes show evidence of differential expression (FDR< 0.05) and 45/72 drug-eQTL interaction genes. This approach highlights that many interactions are being driven by changes in variance of gene expression across the environmental variables rather than necessarily changes in mean expression and therefore eQTL interactions can offer additional information to what is identified through traditional differential expression analysis.

We note that as we have conducted our study on whole blood, some of our observed interactions could be driven by variability in cell type proportions between individuals or as a consequence of cellular subpopulation frequencies being altered by the environmental perturbagen. A limitation of this study is that we lack the complete blood counts to explore this thoroughly. However, we determined B and T cell abundance from FACS data and used CIBERSORT [40] to deconvolute the relative proportions of nine hematopoietic populations from the RNA-seq data to explore this ("Cell counts"). While the number of significant eQTL interactions is reduced after correcting for cell populations, particularly for drug-eQTL interactions after correcting for the FACS proportions, the interaction effects remain very highly correlated suggesting that the majority of these effects are not being altered by these cell compositions. However, further studies will be required to determine if cytokine shifts are altering cellular populations that were not detected by these actual or inferred cell counts, or the principal components that we included in our analyses. For future studies, it will be informative to quantify a broader range of relative cell types and data from single-cell technologies may be particularly powerful for determining cell type-specific eQTLs [47] and their interactions.

We speculate that drug-eQTL interactions might offer an alternative pharmacogenetic strategy to assess drug response. For many biologic medications, predictive pharmacogenetics through typical association studies has been challenging; for example, studies trying to define genetic or transcriptomic biomarkers of anti-TNF response have not been successful [48, 49]. An eQTL interaction approach can be used to define a genotype-aware score reflecting the biological activity

that a medication is having upon an individual, given their allelic combination of multiple genetic markers. For example, we can define a simple anti-IL-6 exposure score based on 7 anti-IL-6 eQTL interactions with a more stringent FDR (FDR < 0.1). The rationale for this drug exposure score is that the expression of a drug-eQTL interaction gene will reflect the effectiveness of the drug in the individual but will be dependent on the genotype of the eQTL interaction SNP. The score is therefore based on assessing whether the expression of the eQTL target gene was more consistent with the drug exposed or the unexposed state for the corresponding interaction SNP genotype. Unsurprisingly, we found a difference in drug exposure score between the unexposed and exposed samples (Additional file 1: Figure S23) ($r_s = 0.40$, $p = 2.1 \times 10^{-16}$); these differences reflect the fact that the eQTLs were themselves identified by examining samples with and without drug exposure. However, while we did not utilize the administered drug dose to identify drug-eQTL interactions, we observed a significant correlation between drug dose (10, 50, or 200 mg) and drug exposure score ($r_s = 0.16$, $p = 0.02$) in the drug-exposed samples (Additional file 1: Figure S24). A simple eQTL interaction score may therefore have the potential to stratify individuals when assessing response to a medication, for example, those with a higher drug exposure score may have a better response to treatment. Similarly, this score could be correlated with adverse effects to capture informative gene expression signatures.

We do not find an association between anti-IL-6 exposure and IFN status and only eight of the cytokine eQTL interactions overlap. Arguably an anti-cytokine therapeutic that is truly effective in SLE might be expected to reduce IFN levels, given how central IFN is to SLE pathogenesis [50]. However, we note a limitation of this study is that the drug itself did not achieve its primary efficacy endpoint of improving SLE outcomes. Hence, while the drug exposure score for this study tracked with the biological effect of the drug (reducing free IL-6 protein levels), it might not be useful for SLE specifically. However, such a scoring system could be implemented easily in most phase III trials for a broad range of therapeutics, where the numbers of samples are far in excess of this phase II trial, ensuring better powered and more accurate eQTL-interaction mapping.

Conclusions

We devised a framework for identifying in vivo eQTL interactions with therapeutically relevant variables, exploiting repeat measurements from a clinical trial. We have applied this approach to demonstrate how downstream regulatory effects of cytokine biology can be elucidated. This same approach can be applied to a wide range of other clinically important cytokines, their antagonists, or indeed other targeted biologic therapies. We speculate that this approach might even be applied to the presence or absence of disease, or disease activity. However, given the multifaceted nature of disease effects, interpreting an eQTL interaction in that context might be more challenging. Modern clinical cohorts and clinical trial data sets with RNA-seq data that has been collected will make this approach easily applicable on a wide scale.

Methods

Study design

The objectives of this study were to map eQTLs in a cohort of lupus patients and identify eQTL interactions with environmental perturbations such as drug treatment to shed light on drug and disease mechanisms. SLE patients were recruited to a phase II clinical trial to test the efficacy and safety of an IL-6 monoclonal antibody (PF-04236921). The patient population recruited to this trial have been detailed extensively by Wallace et al. [8]. One hundred eighty-three patients (forming a multi-ethnic cohort) were randomized to receive three doses of drug (10, 50, or 200 mg) or placebo at three time points during the trial (weeks 0, 8, and 16). Table 1 summarizes the number of patients and samples available.

RNA sequencing

We collected peripheral venous blood samples in PAXgene Blood RNA tubes (PreAnalytiX GmbH, BD Biosciences) for high-depth RNA-seq profiling at 0, 12, and 24 weeks. We extracted total RNA from blood samples using the PAXgene Blood RNA kit (Qiagen) at a contract lab using a customized automation method. We assessed the yield and quality of the isolated RNA using Quant-iT™ RiboGreen® RNA Assay Kit (Thermo Fisher Scientific) and Agilent 2100 Bioanalyzer (Agilent Technologies), respectively. Following quality assessment, we processed an aliquot of 500–1000 ng of each RNA with a GlobinClear-Human kit (Thermo Fisher Scientific) to remove globin mRNA. We then converted RNA samples to cDNA libraries using TruSeq RNA Sample Prep Kit

Table 1 Summary of patients and samples available for each data type. Where relevant, the number of patients/samples remaining after quality control (QC) is displayed in brackets

Data	Patients (post-QC)	Samples (post-QC)
Study design	183	549
RNA sequencing	180 (180)	468 (464)
Genotyping	160 (159)	
eQTL analysis	157	379
IFN status	157	376
Free IL-6 protein levels	145	311
T and B cell counts	152	320

v2 (Illumina) and sequenced using Illumina HiSeq 2000 sequencers. We generated an average of 40 M 100 bp pair-end reads per sample for downstream analysis.

We successfully obtained 468 RNA-seq profiles from 180 patients. We aligned reads to the reference genome (GENCODE [51] release 19) and quantified gene expression using Subread [52] and featureCounts [53] respectively. We included genes with at least 10 reads (CPM > 0.38) in at least 32 samples (minimum number of patients with both unexposed and exposed RNA-seq assays in a drug group) prior to normalization. Following QC, we removed four samples as outliers. We then normalized 20,253 transcripts using the trimmed mean of *M*-values method and the edgeR R package [54]. Expression levels are presented as $\log_2(\text{cpm} + 1)$ and available through figshare [34].

Genotyping

We genotyped 160 individuals across 964,193 variants genome-wide with the Illumina HumanOmniExpressExome-8v1.2 beadchip. We removed SNPs if they deviated from Hardy-Weinberg Equilibrium (HWE) ($p < 1 \times 10^{-7}$), had a minor allele frequency < 5%, missingness > 2%, or a heterozygosity rate greater than 3 standard deviations from the mean (PLINK [55, 56]). For mapping eQTLs, we removed SNPs on the Y chromosome. Following QC, we used 608,017 variants for further analysis. We removed one sample with high missingness and outlying heterozygosity rate from further analysis.

Imputation

We pre-phased the genotypes with SHAPEIT v2 [57]. We imputed missing genotypes and untyped SNPs using Impute2 [58] in 5 Mb chunks against the 1000 Genomes Phase 3 [59] reference panel. To ensure only high-quality genotypes, and to avoid artifacts that can be induced by imputation uncertainty, we removed SNPs with an info score < 1, MAF < 0.05, or HWE $p < 1 \times 10^{-7}$ leaving 1,595,793 SNPs for further analysis.

Interferon status

We classified the interferon (IFN) status of each sample at each time point from the expression of 11 IFN response genes (*HERC5*, *IFI27*, *IRF7*, *ISG15*, *LY6E*, *MX1*, *OAS2*, *OAS3*, *RSAD2*, *USP18*, *GBP5*) using TaqMan Low Density Arrays. These 11 genes were selected by identifying transcripts for which there was both a measureable response to IFN treatment in vitro, as well as differential expression (reduction in expression level) between baseline and visits with clinical improvement in the BOLD study [35]. There is no consensus set of genes to determine the IFN status of SLE patients but these 11 genes do overlap with other published gene sets. For example, 4/11 genes are also used in the 7-gene set defined by McBride et al. [60] and 9/11 genes overlap with the 21-gene set defined by Yao et al. [61].

The first principal component of the expression of the 11-gene set captured 91.7% of the variation (Additional file 1: Figure S25). The distribution of this first principal component is nearly bimodal with good separation (Fig. 2a) and we classified samples as high or low IFN based on this first principal component score. In our dataset, we see excellent correlations ($r_s = 0.86–0.98$) between the real-time PCR expression and the RNA-seq expression for these 11 genes (Additional file 1: Figure S26). The first PC of the IFN signature of RNA-seq data is also strongly correlated with the first PC of the IFN signature of real-time PCR ($r_s = 0.96$, Additional file 1: Figure S27). IFN status was available for 376 samples from 157 subjects.

Drug exposure

Samples were assigned as unexposed (placebo or week 0 samples) or drug exposed (week 12 and week 24 samples in the drug groups).

Free IL-6 protein levels

We determined free IL-6 protein levels from serum using a commercial sandwich ELISA selected for binding only free IL-6. The assay was validated according to FDA biomarker and fit-for purpose guidelines. Free IL-6 protein levels were available for 311 samples from 145 subjects. Since the distribution of IL-6 levels was highly skewed, we ranked samples in order of IL-6 protein levels and included in the model to identify drug-eQTL interactions.

Statistical analysis

eQTL and interaction analysis

In total, 157 patients (with 379 RNA-seq samples) had good quality gene expression and genotyping data for eQTL analysis. All statistical analyses were carried out in R [62].

We defined a *cis* eQTL as the SNP within 250 kb upstream of the GENCODE [51] transcription start site of the gene or 250 kb downstream of the transcription end site. We first applied a linear model for the first available time point (week 0 sample for $n = 152$, week 12 sample for $n = 5$) to identify each eQTL using the first 25 principal components of gene expression and the first 5 principal components of genotyping as covariates.

To select the number of gene expression principal components to include, we counted the number of eQTL genes identified after incrementally increasing the number of principal components accounted for in the model from 0 to 50 by increments of five (Additional file 1: Figure S28). We selected 25 principal components of gene expression to maximize the number of eQTL genes detected while minimizing the number of

principal components we corrected for. We included 5 principal components of genotyping to account for the heterogeneity in ethnicity in our cohort (Additional file 1: Figure S29).

SNPs were encoded as 0, 1, and 2 with respect to the number of copies of the minor allele. To adjust for multiple testing during eQTL discovery, we used a stringent Bonferroni corrected p value threshold of 8.5×10^{-9} (0.05/5,872,001 tests). The Bonferroni adjustment assumes independence among the tests, and we therefore note that it is a conservative multiple comparisons adjustment.

To map eQTLs using multiple samples for each individual, we applied a random intercept linear mixed model using the first 25 principal components of gene expression and the first 5 principal components of genotyping as covariates and patient as a random effect:

$$E_{i,j} = \theta + \beta_{\text{geno}} \cdot g_j + (\kappa_i | j) + \sum_{l=1}^{25} \phi_l \cdot pc_{i,l} + \sum_{m=1}^{5} \gamma_m \cdot pc_{j,m}$$

where $E_{i,j}$ is gene expression for the ith sample from the jth subject, θ is the intercept, β_{geno} is the effect (eQTL) of the genotype for individual j (g_j), $(\kappa_i|j)$ is the random effect for the ith sample from the jth subject, ϕ_l is the effect of principal component l of gene expression for sample i ($pc_{i,l}$), and γ_m is the effect of principal component m of genotyping for subject j ($pc_{j,m}$).

We fitted the linear mixed models using the lme4 R package [63]. We assumed covariance between samples from the same individual, but did not assume any structure in this covariance.

We used the most significant SNP (with $p < 8.5 \times 10^{-9}$) from the 4818 identified eQTL genes to explore eQTL interactions. For each environmental interaction analysis, we further filtered these eQTLs to include only those with at least two individuals homozygous for the minor allele of the SNP being tested in each of the environmental factor groups. For example, we required two of these individuals in each of the drug exposed and drug unexposed groups. To identify eQTL interactions, we added an additional covariate to the model for example drug exposure, and an interaction term between this covariate and the genotype of the SNP:

$$E_{i,j} = \theta + \beta_{\text{geno}} \cdot g_j + (\kappa_i | j) + \sum_{l=1}^{25} \phi_l \cdot pc_{i,l} + \sum_{m=1}^{5} \gamma_m \cdot pc_{j,m} + \beta_{\text{drug}} \cdot d_i + \beta_x \cdot d_i \cdot g_j$$

where $E_{i,j}$ is gene expression for the ith sample from the jth subject, θ is the intercept, β_{geno} is the effect (eQTL) of the genotype for individual j (g_j), $(\kappa_i|j)$ is the random effect for the ith sample from the jth subject, ϕ_l is the effect of principal component l of gene expression for sample i ($pc_{i,l}$), γ_m is the effect of principal component m of genotyping for subject j ($pc_{j,m}$), β_{drug} is the effect (differential gene expression) of drug for sample i (d_i), and β_x is the effect of the drug genotype interaction ($d_i \bullet g_j$).

We determined the significance of the interaction term with a likelihood ratio test.

To rigorously confirm the relative enrichment of eQTL interactions, we shuffled the interaction covariate (for example drug exposure) 1000 times and calculated the number of significant interactions observed in each permutation. Our primary goal for the permutation analysis was to retain the main eQTL effect while examining only the effect of the environmental factor on the interaction. In this study, the main purpose of the covariates included in the model is to ensure the main eQTL effect is found. For IFN high/low status, we shuffled across all samples. For drug interaction permutation analysis, we maintained the number of individuals in the drug group and the number of samples with exposure to drug. We calculated a q value for each interaction using the q value package [36]. Additional file 1: Figure S30 shows the observed versus the expected p values for the interaction analyses.

The expression of the majority of genes followed a normal distribution (Additional file 1: Figure S31) but to assess whether non-normality could be causing an inflation of our test statistic, we repeated the identification of eQTLs and eQTL interactions following the standard normal transformation. We transformed the expression values of each gene to their respective quantiles of a normal distribution using the qqnorm function in R, breaking any ties (for example expression levels of zero in some individuals) randomly.

Concordance with an eQTL study in healthy individuals

In the SLE cohort, we classified 4818 *cis* eQTL genes ($p < 8.5 \times 10^{-9}$). The z-score for the most associated SNP for each of these genes was compared to the z-score from a previously published eQTL dataset from whole blood from 2166 healthy individuals [11]. 4113/4818 SNP-gene pairs (85.4%) were also reported in the BIOS dataset (FDR < 0.05). After removing 301 SNPs, which could not be mapped to a strand, 3770/3812 (98.9%) had a z-score (eQTL effect) in a consistent direction.

Magnifiers and dampeners

An eQTL interaction can either magnify or dampen the original eQTL effect. We multiplied the interaction z-score by the sign of the original eQTL effect (genotype beta) and defined magnifiers as interactions with an

adjusted z-score > 0 and dampeners as interactions with an adjusted z-score < 0.

Differential gene expression analysis

To identify differentially expressed genes following drug exposure (unexposed or exposed), we applied a random intercept linear mixed model with patient as a random effect. We calculated a q value using the q value package [36].

Drug exposure score

We assigned a drug exposure score to each sample. We calculated a score for each gene (see equation below) and then averaged across the seven drug-eQTL genes (FDR < 0.1) to give the final drug exposure score.

$$\text{Drug exposure score for gene} = \frac{1}{2}\left(\frac{G-G_{\text{Unexp}}}{\text{SE}}\right)^2 - \frac{1}{2}\left(\frac{G-G_{\text{Exp}}}{SE}\right)^2$$

where G is gene expression for a given sample, G_{Unexp} is predicted mean gene expression for unexposed samples of the relevant SNP genotype, G_{Exp} is predicted mean gene expression for exposed samples of the relevant SNP genotype, and SE is standard error for the intercept term of the model (unexposed expression for genotype 0).

HOMER analysis for transcription factor binding motif enrichment

We used the HOMER software suite [37] to look for enrichment of transcription factor binding motifs in the 210 IFN-eQTL interactions (FDR < 0.2) and the 72 drug-eQTL interactions (FDR < 0.2). Each eQTL interaction was identified using the most highly associated SNP for that eQTL. However, as this SNP is not necessarily the functional SNP, we additionally considered all those with an $r^2 \geq 0.8$ in the 1000 Genomes European population [59] within the *cis* eQTL window. We defined our motif search window as 20 bp on either side of each SNP (i.e., 41 bp wide).

For each environmental factor, we divided the eQTL interactions into magnifiers or dampeners and conducted two separate HOMER analyses: the proportion of magnifying eQTL interaction sequences with a motif compared to the proportion of dampening interaction sequences with a motif and vice versa. HOMER reported the transcription factor motifs that were significantly enriched in one category of interactions relative to the other. Motifs were plotted using the SeqLogo R library [64].

We determined permutation p values for enrichment of the ISRE and IRF4 transcription factor binding sites as follows. For ISRE, the motif is interrupted by interaction SNPs (or SNPs in LD) corresponding to 11 magnifying genes and 1 dampening gene. We permuted which genes were labeled as magnifiers or dampeners 100,000 times and counted the number of genes in each category with an ISRE motif interrupted. We found 1855 occurrences from 100,000 trials with at least 11 magnifying genes ($p < 0.019$). For IRF4, the motif is interrupted by SNPs corresponding to 9 magnifying genes and 4 dampening genes. Using the same permutation approach, we found 5801 occurrences from 100,000 trials with at least 9 magnifying genes ($p < 0.058$).

Cell counts

We collected 4 ml whole blood in sodium heparin vacutainers for cytometry analysis at weeks 0, 12, and 24. Samples were subjected to flow cytometry for T cell and B cell immunophenotyping (Additional file 1: Figure S32). We counted T (CD3+) and B (CD19+) cells as a percentage of lymphocytes (CD45+, SSC-small) because of the abnormal distribution of lymphocytes observed in SLE [65]. These counts are therefore inversely correlated ($r_s = -0.65$, Additional file 1: Figure S33). FACS data were available for 320 samples from 152 subjects.

We used CIBERSORT [40] to deconvolute proportions of cell types from the RNA-seq data. We used the LM22 database from CIBERSORT which contains cell signatures for 22 cell types and grouped these into nine representative cell types (eosinophils, neutrophils, B cells, T cells, natural killer cells, macrophages, dendritic cells, mast cells, and monocytes).

Additional files

Additional file 1: Supplementary Figure S1-S33. (PDF 8033 kb)
Additional file 2: Supplementary Tables S1-S6. (XLSX 6624 kb)
Additional file 3: Review history. (DOCX 841 kb)
Additional file 4: Supplementary Table S7. (PDF 148 kb)

Review history

The review history is available as Additional file 3.

Funding

This work is supported in part by funding from the National Institutes of Health (U01GM092691, UH2AR067677, U19AI111224 (SR)), the Doris Duke Charitable Foundation Grant #2013097, the Ruth L. Kirschstein National Research Service Award (F31AR070582) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (KS) and the Rheumatology Research Foundation Tobé and Stephen E. Malawista, MD Endowment in Academic Rheumatology (D.A.R). This work is also supported by unrestricted funding from Pfizer, Inc.

Availability of data and materials

Data are available in the supplementary tables and at http://baohongz.github.io/Lupus_eQTL [32]. Raw RNA-seq data are available through GEO (GSE116006) [66]. We have deposited the genotyping data in NCBI's dbGaP (phs001702.v1.p1) [67]. Normalized gene expression data and eQTL summary statistics are available through figshare [34].

Authors' contributions

The project was conceived and designed by EED, MSV, BZ, and SR. Statistical analysis was conducted by EED, TA, MG-A, KS, H-JW, YL, and CS. Molecular

data was obtained, organized and analyzed by YZ, SP, DvS, JSB, NB, MSV, BZ, and DAR. The initial manuscript was written by EED and SR. All authors edited and approved the manuscript.

Ethics approval and consent to participate

The protocol (2014P002477) for analysis and data sharing was approved by the institutional review board of Brigham and Women's Hospital subject to applicable laws and regulations and ethical principles consistent with the Declaration of Helsinki. The names of the ethics committees that approved the subject recruitment at each site are provided in Additional file 4: Table S7. All subjects gave written informed consent.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA 02115, USA. [2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. [3]Partners Center for Personalized Genetic Medicine, Boston, MA 02115, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. [6]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA. [7]Division of Rheumatology, Allergy, Immunology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. [8]Pfizer Inc., Cambridge, MA 02139, USA. [9]Pfizer New Haven Clinical Research Unit, New Haven, CT 06511, USA. [10]Biogen, Cambridge, MA 02142, USA. [11]Faculty of Medical and Human Sciences, University of Manchester, M13 9PL, Manchester, UK. [12]Harvard New Research Building, 77 Avenue Louis Pasteur, Suite 250D, Boston, MA 02446, USA.

Received: 29 January 2018 Accepted: 5 October 2018
Published online: 19 October 2018

References

1. Rider P, Carmi Y, Cohen I. Biologics for targeting inflammatory cytokines, clinical uses, and limitations. Int J Cell Biol. 2016;2016:iv.
2. Cessak G, Kuzawińska O, Burda A, Lis K, Wojnar M, Mirowska-Guzel D, et al. TNF inhibitors - mechanisms of action, approved and off-label indications. Pharmacol Reports. 2014;66:836–44.
3. Tanaka Y, Mola EM. IL-6 targeting compared to TNF targeting in rheumatoid arthritis: studies of olokizumab, sarilumab and sirukumab. Ann Rheum Dis. 2014;73:1595–7.
4. Stone JH, Tuckwell K, Dimonaco S, Klearman M, Aringer M, Blockmans D, et al. Trial of tocilizumab in giant-cell arteritis. N Engl J Med. 2017;377:317–28. https://doi.org/10.1056/NEJMoa1613849.
5. Van Rhee F, Wong RS, Munshi N, Rossi JF, Ke XY, Fosså A, et al. Siltuximab for multicentric Castleman's disease: a randomised, double-blind, placebo-controlled trial. Lancet Oncol. 2014;15:966–74.
6. Tackey E, Lipsky PE, Illei GG. Rationale for interleukin-6 blockade in systemic lupus erythematosus. Lupus. 2004;13:339–43.
7. Illei GG, Shirota Y, Yarboro CH, Daruwalla J, Tackey E, Takada K, et al. Tocilizumab in systemic lupus erythematosus: data on safety, preliminary efficacy, and impact on circulating plasma cells from an open-label phase I dosage-escalation study. Arthritis Rheum. 2010;62:542–52.
8. Wallace DJ, Strand V, Merrill JT, Popa S, Spindler AJ, Eimon A, et al. Efficacy and safety of an interleukin 6 monoclonal antibody for the treatment of systemic lupus erythematosus: a phase II dose-ranging randomised controlled trial. Ann Rheum Dis. 2017;76:534–42. https://doi.org/10.1136/annrheumdis-2016-209668.
9. Bronson PG, Chaivorapol C, Ortmann W, Behrens TW, Graham RR. The genetics of type I interferon in systemic lupus erythematosus. Curr Opin Immunol. 2012;24:530–7. https://doi.org/10.1016/j.coi.2012.07.008.
10. Ridker PM, Everett BM, Thuren T, MacFadyen JG, Chang WH, Ballantyne C, et al. Antiinflammatory therapy with Canakinumab for atherosclerotic disease. N Engl J Med. 2017;377:119–1131. https://doi.org/10.1056/NEJMoa1707914.
11. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017;49:139–45. https://doi.org/10.1038/ng.3737.
12. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343:1246949. https://doi.org/10.1126/science.1246949.
13. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012;44:502–10. https://doi.org/10.1038/ng.2205.
14. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014;344:519–23.
15. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET, et al. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60. https://doi.org/10.1126/science.1262110.
16. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011;7:e1002003. https://doi.org/10.1371/journal.pgen.1002003.
17. Kukurba KR, Parsana P, Balliu B, Smith KS, Zappala Z, Knowles DA, et al. Impact of the X chromosome and sex on regulatory variation. Genome Res. 2016;26:768–77.
18. Buil A, Brown AA, Lappalainen T, Viñuela A, Davies MN, Zheng H-F, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. Nat Genet. 2015;47:88–91. https://doi.org/10.1038/ng.3162.
19. Maranville JC, Luca F, Stephens M, Di Rienzo A. Mapping gene-environment interactions at regulatory polymorphisms: insights into mechanisms of phenotypic variation. Transcription. 2012;3:56–62. https://doi.org/10.4161/trns.19497.
20. Idaghdour Y, Awadalla P. Exploiting gene expression variation to capture gene-environment interactions for disease. Front Genet. 2013;4:1–7.
21. Idaghdour Y, Quinlan J, Goulet J-P, Berghout J, Gbeha E, Bruat V, et al. Evidence for additive and interaction effects of host genotype and infection in malaria. Proc Natl Acad Sci U S A. 2012;109:16786–93. https://doi.org/10.1073/pnas.1204945109.
22. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, et al. Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet. 2010;42:62–7. https://doi.org/10.1038/ng.495.
23. Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, et al. Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. PLoS Genet. 2016;12:e1005908. https://doi.org/10.1371/journal.pgen.1005908.
24. Li Y, Álvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet. 2006;2:2155–61.
25. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. PLoS Biol. 2008;6:810–24.
26. Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, Baxter S, et al. Interactions between glucocorticoid treatment and Cis-regulatory polymorphisms contribute to cellular response phenotypes. PLoS Genet. 2011;7:e1002162.
27. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012;109:1204–9.
28. Smirnov D, Morley M, Shin E. Genetic analysis of radiation-induced changes in human gene expression. Nature. 2009;459:587–91. https://doi.org/10.1038/nature07940.Genetic.
29. Walsh AM, Whitaker JW, Huang CC, Cherkas Y, Lamberth SL, Brodmerkel C, et al. Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. Genome Biol. 2016;17:79. https://doi.org/10.1186/s13059-016-0948-6.
30. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Ortmann WA, Espe KJ, et al. Interferon-inducible gene expression signature in peripheral blood

cells of patients with severe lupus. Proc Natl Acad Sci U S A. 2003;100:2610–5.
31. Bennett L, Palucka AK, Arce E, Cantrell V, Borvak J, Banchereau J, et al. Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. J Exp Med. 2003;197:711–23.
32. Zhang B. Lupus_eQTL. 2018. http://baohongz.github.io/Lupus_eQTL.
33. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of Cis regulatory variation in diverse human populations. PLoS Genet. 2012;8:e1002639.
34. Davenport EE, Amariuta T, Gutierrez-Arcelus M, Slowikowski K, Westra H-J, Luo Y, et al. Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial datasets. figshare. 2018. https://doi.org/10.6084/m9.figshare.7105202.
35. Hill AA, Immermann FW, Zhang Y, Reddy PS, Zhou T, O'Toole M, et al. FRI0003 determination of interferon (IFN) signatures for SLE patients may be critical for optimal treatment selection but depends on the genes chosen: report from the bold (biomarkers of lupus disease) study. Ann Rheum Dis. 2013;72:A369–70. https://doi.org/10.1136/annrheumdis-2013-eular.1131.
36. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100:9440–5. https://doi.org/10.1073/pnas.1530509100.
37. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89. https://doi.org/10.1016/j.molcel.2010.05.004.
38. Au-Yeung N, Mandhana R, Horvath CM. Transcriptional regulation by STAT1 and STAT2 in the interferon JAK-STAT pathway. Jak-Stat. 2013;2:e23931. https://doi.org/10.4161/jkst.23931.
39. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43. https://doi.org/10.1038/ng.2756.
40. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
41. Koch S, Mousset S, Graser A, Reppert S, Übel C, Reinhardt C, et al. IL-6 activated integrated BATF/IRF4 functions in lymphocytes are T-bet-independent and reversed by subcutaneous immunotherapy. Sci Rep. 2013;3:1754. https://doi.org/10.1038/srep01754.
42. Lech M, Weidenbusch M, Kulkarni OP, Ryu M, Darisipudi MN, Susanti HE, et al. IRF4 deficiency abrogates lupus nephritis despite enhancing systemic cytokine production. J Am Soc Nephrol. 2011;22:1443–52. https://doi.org/10.1681/ASN.2010121260.
43. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506:376–81. https://doi.org/10.1038/nature12873.
44. Xu WD, Pan HF, Ye DQ, Xu Y. Targeting IRF4 in autoimmune diseases. Autoimmun Rev. 2012;11:918–24. https://doi.org/10.1016/j.autrev.2012.08.011.
45. Furie R, Merrill J, Werth V, Khamashta M, Kalunian K, Brohawn P, et al. Anifrolumab, an anti-interferon alpha receptor monoclonal antibody, in moderate to severe systemic lupus erythematosus (SLE). Arthritis Rheumatol. 2017;69:376–86.
46. Lappalainen T, Sammeth M, Friedländer MR, 't hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11. https://doi.org/10.1038/nature12531.
47. Van Der Wijst MGP, Brugge H, De Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50:493–7. https://doi.org/10.1038/s41588-018-0089-9.
48. Cui J, Stahl EA, Saevarsdottir S, Miceli C, Diogo D, Trynka G, et al. Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. PLoS Genet. 2013;9:e1003394. https://doi.org/10.1371/journal.pgen.1003394.
49. Cui J, Diogo D, Stahl EA, Canhao H, Mariette X, Greenberg MPHJD, et al. The role of rare protein-coding variants to anti-TNF treatment response in rheumatoid arthritis. Arthritis Rheumatol. 2017;69:735–41.
50. Davis LS, Hutcheson J, Mohan C. The role of cytokines in the pathogenesis and treatment of systemic lupus erythematosus. J Interf Cytokine Res. 2011;31:781–9. https://doi.org/10.1089/jir.2011.0047.
51. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74. https://doi.org/10.1101/gr.135350.111.
52. Liao Y, Smyth GK, Shi W. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41:e108. https://doi.org/10.1093/nar/gkt214.
53. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
54. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. https://doi.org/10.1093/bioinformatics/btp616.
55. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7. https://doi.org/10.1186/s13742-015-0047-8.
56. Purcell S, Chang C. PLINK 1.9. https://www.cog-genomics.org/plink2.
57. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9:179–81. https://doi.org/10.1038/nmeth.1785.
58. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529. https://doi.org/10.1371/journal.pgen.1000529.
59. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. https://doi.org/10.1038/nature15393.
60. McBride JM, Jiang J, Abbas AR, Morimoto A, Li J, Macluca R, et al. Safety and pharmacodynamics of rontalizumab in patients with systemic lupus erythematosus: results of a phase I, placebo-controlled, double-blind, dose-escalation study. Arthritis Rheum. 2012;64:3666–76.
61. Yao Y, Higgs BW, Morehouse C, de Los RM, Trigona W, Brohawn P, et al. Development of potential pharmacodynamic and diagnostic markers for anti-IFN-α monoclonal antibody trials in systemic lupus erythematosus. Hum Genomics Proteomics. 2009;2009:374312. https://doi.org/10.4061/2009/374312.
62. R Core Team. R: a language and environment for statistical computing. 2015. https://www.r-project.org.
63. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48. https://doi.org/10.18637/jss.v067.i01.
64. Bembom O. seqLogo: sequence logos for DNA sequence alignments. 2016.
65. Shirota Y, Yarboro C, Fischer R, Pham T, Lipsky P, Illei GG. Impact of anti-interleukin-6 receptor blockade on circulating T and B cell subsets in patients with systemic lupus erythematosus. Ann Rheum Dis. 2013;72:118–28.
66. Davenport EE, Amariuta T, Gutierrez-Arcelus M, Slowikowski K, Westra H-J, Luo Y, et al. Discovering in vivo cytokine eQTL interactions from a lupus clinical trial. GEO. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116006.
67. Raychaudhuri S. eQTL Interactions in Lupus Patients. dbGaP. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001702.v1.p1.

Published in final edited form as:
Nat Genet. 2018 April ; 50(4): 493–497. doi:10.1038/s41588-018-0089-9.

Single-cell RNA sequencing identifies cell type-specific *cis*-eQTLs and co-expression QTLs

Monique G.P. van der Wijst, **Harm Brugge**[#], **Dylan H. de Vries**[#], **Patrick Deelen**, **Morris A. Swertz**, **LifeLines Cohort Study**, **BIOS Consortium**, and **Lude Franke**[*]
Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[#] These authors contributed equally to this work.

Genome-wide association studies have identified thousands of genetic variants that are associated with disease.1 Most of these variants have small effect sizes, but their downstream expression effects, so-called expression quantitative trait loci (eQTLs), are often large2 and cell type-specific3–5. To identify these cell type-specific eQTLs using an unbiased approach, we used single-cell RNA sequencing (scRNA-seq) to generate expression profiles of ~25,000 peripheral blood mononuclear cells (PBMCs) from 45 donors. We identified previously reported *cis*-eQTLs, but also identified new cell type-specific *cis*-eQTLs. Finally, we generated personalized co-expression networks, and identified genetic variants that significantly alter co-expression relationships (which we termed 'co-expression QTLs'). Single-cell eQTL analysis thus allows for the identification of genetic variants that impact regulatory networks.

Previously, purified cell types4,6–8 or deconvolution methods9,10 have been used to identify cell type-specific eQTLs. However, these methods are biased towards specific cell types, or are of limited use for less abundant cell types and dependent on accurately defined marker genes.11 In contrast, scRNA-seq can be used to investigate rare cell types12, and thus, enables identification of cell type-specific eQTLs using an unbiased approach. Indeed, proof of concept was previously shown in a study on 15 individuals, where 92 genes were studied in 1,440 cells.13

[*]Correspondence should be addressed to L.F.: lude@ludesign.nl.

Author contribution
MW generated the scRNA-seq data. MW, HB and DV performed bioinformatics and statistical analyses. PD and BIOS consortium performed replication of co-expression QTLs. MW and LF designed the study and wrote the manuscript. MS and the LLD consortium provided biomaterials, genotype data and computational resources. All authors discussed the results and commented on the manuscript.

Competing financial interests
The authors declare no competing financial interests.

Ethics approval and consent to participate
The LifeLines DEEP study was approved by the ethics committee of the University Medical Centre Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent form prior to study enrollment. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Europe PMC Funders Author Manuscripts

Here, we studied cell type-specific effects of genetic variation on genome-wide gene expression by generating scRNA-seq data of ~25,000 PBMCs from 45 donors of the population-based cohort study Lifelines Deep14. After quality control (Online Methods, Suppl. Fig. 1), we first assessed to which extent previously reported *cis*-eQTLs from bulk whole blood, using either 94 DeepSAGE samples 15 (a 3'-end oriented RNA-sequencing strategy similar to our scRNA-seq approach) or 2,116 RNA-seq11 samples, also show significant effects in the scRNA-seq dataset. For this analysis, we treated the scRNA-seq data as being bulk PBMCs (by averaging expression levels of all cells per gene per sample, referred to as 'bulk-like PBMCs'). We detected 50 and 311 significant *cis*-eQTLs (gene-level false-discovery rate (FDR) of 0.05) that were previously reported in the DeepSAGE15 and RNA-seq11 study, respectively (Fig. 1a, Suppl. Table 1). Although only a small proportion (8% and 1%) of previously reported *cis*-eQTLs were significant in our scRNA-seq analysis, 96% and 90.4% had identical allelic directions as in the DeepSAGE15 and RNA-seq11 study, respectively, indicating that these *cis*-eQTLs reflect similar regulatory effects. The few discordant eQTLs may reflect the slightly different sample composition of both datasets (PBMCs versus whole blood) and the relatively few sequence reads targeting the 3'-end of genes in the bulk RNA-seq dataset.

We subsequently performed a genome-wide *cis*-eQTL discovery analysis on the bulk-like PBMCs. Separate *cis*-eQTL analyses were conducted on each of the identified major cell types (cell type classification was performed using Seurat16, Suppl. Fig. 2a, 2b) by averaging the normalized gene expression of all cells per cell type, gene and donor. In total, 379 unique top *cis*-eQTLs were identified, reflecting 287 unique eQTL genes (gene-level FDR of 0.05) (Table 1), as sometimes in different cell types different SNPs showed the most significant association for an eQTL gene. While 331 (reflecting 249 unique *cis*-eQTL genes) of these 379 *cis*-eQTLs were significant in the bulk-like PBMC eQTL analysis, 48 *cis*-eQTLs (reflecting 38 unique *cis*-eQTL genes) were only detected in specific cell types (i.e. 'cell type-dependent' eQTLs, Suppl. Table 2).

We subsequently attempted to replicate these eQTLs. For the 249 eQTL genes found in the bulk-like PBMC analysis, 233 *cis*-eQTLs were testable and 181 (78%) were associated with the same SNP (90.1% shared allelic direction, Suppl. Table 2) in the whole blood RNA-seq eQTL data set11. For the 48 cell type-dependent *cis*-eQTLs, 29 (60%) replicated in the RNA-seq dataset11. This lower percentage suggests that in bulk RNA-seq datasets cell type-dependent eQTLs might become too diluted, resulting in low statistical power to recover these. While this most likely happens for rare cell types, we also observed this in common cell types. For instance, in the most abundant cell type ($CD4^+$ T cells), rs2272245 significantly affects the expression of the *TSPAN13* gene in *cis* ($P = 2.21 \times 10^{-6}$). However, this effect was not significant in the bulk-like PBMCs (P = 0.88), because *TSPAN13* is lowly expressed in CD4+ T-cells, whereas it is highly expressed in dendritic cells (DCs) where it did not show a *cis*-eQTL effect (Fig. 1b). *Cis*-eQTLs might also be missed in bulk data, because they might show opposite allelic effects across different cell types. We could not study this in detail, due to lack of power given the sample size and limited number of cells for rare cell types (Suppl. Fig. 2c). Nevertheless, in $CD4^+$ T cells, the A allele of rs4804315 significantly decreased expression of *ZNF414* in *cis* ($P = 6.09 \times 10^{-6}$), whereas in natural killer (NK) cells this allele increased expression of *ZNF414* at nominal significance (P =

0.0339) (Fig. 1b). However, it cannot be excluded that specifically in NK cells, the effect of rs4804315 on *ZNF414* expression is the result of a residual effect on *ZNF414* expression of a second, independent variant.

Since some *cis*-eQTLs did not replicate in whole blood bulk RNA-seq data, we subsequently investigated eQTL datasets of purified cell types. Indeed, 3 out of 19 remaining cell type-dependent *cis*-eQTLs were detected (each with consistent allelic direction) in purified eQTL datasets of the Blueprint consortium (naïve $CD4^+$ T cells and $CD14^+$ monocytes)17 or Kasela et al. ($CD4^+$ and $CD8^+$ T cells)6 (Suppl. Table 3). Hence, only 16 cell type-dependent *cis*-eQTLs were not identified before using bulk eQTL datasets of blood or purified immune cells. Although some *cis*-eQTLs were only significant in specific cell types, this does not prove cell type-specificity; particularly in less abundant cell types power is lacking to detect many *cis*-eQTLs. Ways to partially overcome this, would be to use methods that consider multiple eQTL datasets together, such as eQTL-BMA18 or Meta-Tissue19. However, these methods are currently computationally too demanding for large scRNA-seq data or do not define the cell type in which the eQTL effect occurs.19,20

A major advantage of using scRNA-seq data is the flexibility by which any cell population of interest can be selected for eQTL analysis. In contrast, when using RNA-seq data of purified cell types, one cannot retrieve data from subcell types anymore. Moreover, while finer differences between subcell types may be detectable using gene expression profiles, it is not always recapitulated by different cell membrane markers, complicating cell sorting. Here, we show the added value of performing eQTL analysis on subcell types using two monocyte subsets: classical (cMonocytes) and non-classical monocytes (ncMonocytes). When plotting Spearman's rank correlation of each top eQTL for the cMonocytes against the ncMonocytes, several examples were revealed that pinpointed the eQTL effect specifically to cMonocytes (Fig. 1c). Two such examples, which were previously identified in RNA-seq data of purified $CD14^+$ monocytes17, are shown in Figure 1d. The scRNA-seq data now allowed us to specifically assign these effects to cMonocytes (Fig. 1d). Despite having lower power for detecting eQTLs in ncMonocytes due to an almost five times lower abundance compared to cMonocytes (Suppl. Fig. 2b), power in the ncMonocytes remains sufficiently high to detect several other significant ncMonocyte *cis*-eQTLs (Fig. 1e, Suppl. Table 2).

Another opportunity of scRNA-seq data is to use it for determining whether genetic variants can alter gene co-expression. Although recently genes and environmental factors altering the effect size of eQTLs ('context-specific eQTLs') have been identified in bulk RNA-seq eQTL datasets11,21, a large sample size was required to ensure sufficient power. In contrast, scRNA-seq data enables generation of co-expression networks on an individual donor basis, which vastly reduces the number of samples required to identify SNPs altering co-expression relationships. This enabled us to study whether SNPs showing *cis*-eQTL effects also affect the co-expression relationship of the *cis*-eQTL genes with other genes, which we further define as 'co-expression QTLs'. We confined our analysis to the most abundant cell type ($CD4^+$ T cells), and calculated the co-expression between individual pairs of genes using Spearman's rank correlation. We restricted the analysis to the 145 *cis*-eQTL genes identified in $CD4^+$ T cells (Table 1), thereby increasing the likelihood of finding co-

expressed genes that are modulated by the same genetic variant. Out of these, 102 genes showed variance in gene expression within each of the 45 donors and were investigated. For two of these genes we identified significant co-expression QTLs: 93 co-expression QTLs were detected for *RPS26* and one for *HLA-B* (P-value ≤ $1.27x10^{-7}$, corresponding to an eQTL-gene level FDR of 0.05). The most significant interaction was found for rs7297175 affecting the co-expression between *RPS26* and *RPL21* (P ≤ $2.70x10^{-16}$) (Fig. 2a, 2b). When using a more liberal FDR of 0.10 (P-value = $4.72x10^{-7}$), we identified significant co-expression QTLs for three eQTL genes (Suppl. Table 4): 13 additional co-expression QTLs were found for *RPS26* and one for *SMDT1.* As a result of co-expression between genes, we cannot rule out that the 106 co-expression QTLs identified for *RPS26* are actually representing just one effect.

To assess the robustness of the identified co-expression QTLs, we tested whether they remained significant after gene expression imputation, which was used to overcome the problem that in scRNA-seq data usually many genes are undetected despite being expressed (i.e. zero-inflated expression). Several computational strategies have been developed to do this.22–24 However, most current methods are either computationally too demanding for large datasets like ours23, or cannot sufficiently impute the 94.1% zero values present in our dataset24. To overcome this, we used MAGIC22, a method that imputes gene expression levels for nearly every gene. To prevent that imputation removes effects of genetic differences between donors or cell types, we performed imputation for each donor separately and again only for $CD4^+$ T cells (see Data availability). In general, imputation worked well, but in some circumstances artifacts were introduced (Suppl. Fig. 3). Therefore, we only used the imputed gene expression data to determine whether the co-expression QTLs, identified prior to imputation, remained significant after imputation (Suppl. Table 4). For the three eQTL genes that were involved in a co-expression QTL, two out of three top co-expression QTLs (rs7297175 affecting the co-expression between *RPS26* and *RPL21,* P = $3.97x10^{-12}$ (Fig. 2c) and rs4147641 affecting the co-expression between *SMDT1* and *RPS3A,* P = $2.57x10^{-4}$) remained after imputation (Suppl. Table 4). Subsequently, we were able to replicate both effects in a whole blood bulk RNA-seq eQTL dataset11 (P = $1.69x10^{-3}$ for *RPS26-RPL21* (Fig. 2d), P = $1.59x10^{-4}$ for *SMDT1-RPS3A*) (Suppl. Table 4). Interestingly, SNP rs7297175, affecting the co-expression between *RPS26* and 106 other genes, is in near perfect linkage disequilibrium with the type I diabetes (T1D) SNP rs1117173925 (r^2 = 0.98). Therefore, the numerous co-expression QTLs for *RPS26* may shed new light on *RPS26* and its link with T1D. This interaction effect was also observed in other cell types (Suppl. Fig. 4), indicating it is not cell type-specific. In addition, various analyses were performed to rule out potential technical confounders (see Online Methods).

The co-expression QTL analysis as outlined above highlights another advantage of scRNA-seq data; with PBMCs from only 45 donors, we could identify effects that would otherwise only become apparent in large-scale (2,116 samples) bulk RNA-seq eQTL datasets11. Due to Simpson's paradox26, it may occur that when looking at all individuals together, the interaction between two genes does not show a correlation, while each of the individuals separately do show a correlation. So, even though the effect may be observed in bulk RNA-seq data, the true correlation will only be revealed using scRNA-seq data.

Europe PMC Funders Author Manuscripts

The eQTL and co-expression QTL analyses performed in this study show the benefit of scRNA-seq data for linking genetic variation to gene expression regulation. In addition to these analyses, we expect scRNA-seq data to offer many other opportunities for selecting cells of interest for eQTL and co-expression QTL analysis. For example, one could use the intercellular variation within scRNA-seq data to group cells along the cell cycle13, along a differentiation path27 or along a response to an environmental stimulus28. By doing so, one might identify eQTLs or co-expression QTLs that are influenced by cell cycle phase, differentiation or environmental status.

In conclusion, this proof of concept study shows the feasibility of using scRNA-seq data for eQTL and gene-gene interaction analysis. The identified eQTLs and co-expression QTLs replicated well with earlier reported whole blood RNA-seq data. Moreover, we extended the list of genes known to be under genetic control or specified the cell type in which the effect is most prominent. Finally, several SNPs were linked to modulation of gene co-expression, implying that gene regulatory networks can be highly personal. We expect that larger single-cell eQTL datasets will enable the identification of many cell type-specific eQTLs and genetic variants that affect regulatory network relationships.

Online methods

Isolation and preparation of PBMCs

Whole blood of 47 donors from the general population Lifelines Deep (LLD) cohort14 was drawn into EDTA-vacutainers (BD). Within 2h, peripheral blood mononuclear cells (PBMCs) were isolated using Cell Preparation Tubes with sodium heparin (BD). For all procedures, PBMCs were kept in RPMI1640 supplemented with 50 μg/mL gentamicin, 2 mM L-glutamine and 1 mM pyruvate. Isolated PBMCs were cryopreserved in RPMI1640 containing 40% FCS and 10% DMSO. Within one month, PBMCs were further processed for scRNA-seq. First, cells were thawed in a 37°C water bath until almost completely thawed, after which the cells were slowly washed in warm medium. After washing, cells were resuspended in medium and incubated for 1h in a 5° slant rack at 37°C in a 5% CO_2 incubator. After this 1h resting period, cells were washed twice in medium supplemented with 0.04% bovine serum albumin. Cells were counted using a haemocytometer and cell viability was assessed by Trypan Blue. Eight, sex-balanced sample pools were prepared each containing 1750 cells/donor from 6 (or 5) donors (10,500 cells).

Single-cell library preparation and sequencing

Single cells were captured using the 10X Chromium controller (10X Genomics) according to the manufacturer's instructions (document CG00026), and as previously described.29 Each sample pool was loaded into a different lane of a 10X chip (Single cell chip kit, 120236). cDNA libraries were generated using the Single Cell 3' Library & Gel Bead kit version 2 (120237) and i7 Multiplex kit (120262) in line with the company's guidelines. These libraries were sequenced using a custom program (27-9-0-138) on 8 lanes of an Illumina HiSeq4000 using a 75bp paired-end kit, per GenomeScan (Leiden, the Netherlands) sequencing guidelines. In total, 28.855 cells were captured and sequenced to an average depth of 74k.

Alignment and initial processing of sequence-data

CellRanger v1.3 software with default settings was used to demultiplex the sequencing data, generate FASTQ files, align the sequencing reads to the hg19 reference genome, filtering of cell and UMI (unique molecular identifier) barcodes, and counting gene expression per cell (see Data availability).

Demuxlet algorithm: demultiplexing samples per lane and doublet detection

Genotypes of the LLD-samples were previously generated14 and were phased using Eagle v2.330 and imputed with the HRC-reference panel31 using the Michigan Imputation Server32. As genotype data of each donor (except 2) was available, we could use the Demuxlet method33 that uses variable SNPs between the pooled individuals to determine which cell belongs to which individual and to identify doublets (two cells encapsulated in a single droplet by the 10X Chromium controller).

To determine how well every genotype matches each cell, a likelihood score was calculated by the formula: $L_c(s) = \prod_{v=1}^{V}\left[\sum_{g=0}^{2}\left\{\prod_{i=1}^{d_{cv}}\left(\sum_{e=0}^{1}\Pr(b_{cvi}|g,e)\right)P_{sv}^{(g)}\right\}\right]$. Here, c is the cell, s is the individual, v are the unique genetic variants (SNPs) found on the reads of the cell, d_{cv} the number of unique reads overlapping with the v^{th} variant from the c^{th} cell. b_{cvi} is the variant-overlapping base call from the i^{th} read, representing reference (R), alternate (A), and other (O) alleles respectively. e_{cvi} is a latent variable indicating whether the base call is correct (0) or not (1) and finally g is the true genotype. This likelihood score was calculated by taking into account the genotype probabilities of a sample at all known SNPs, the variant-overlapping base calls with base quality (Phred quality score) > 15, and a probability that the base was not called correctly, which is fixed at 0.001. In this way, for each pool of cells, the genotype within this pool with the highest likelihood was assigned as the most likely person the cell belonged to.

To identify doublets, likelihoods for a 50/50 ratio of all possible combinations of two genotypes were calculated, similarly as for singlets but now considering two genotypes at the same time. To consider a mix of genotypes from two individuals, the following formula was used:

$$L_c(S_1, S_2, \alpha) = \prod_{v=1}^{V}\left[\sum_{g_1,g_2}\left\{\prod_{i=1}^{d_{cv}}\left(\sum_{e=0}^{1}(1-\alpha)\Pr(b_{cvi}|g_1,e) + \alpha\Pr(b_{cvi}|g_2,e)\right)P_{sv}^{(g_1)}P_{sv}^{(g_2)}\right\}\right].$$

Here, s_1 and s_2 are the two individuals, g_1 and g_2 the corresponding true genotypes and α is the expected proportion of the SNPs in every cell for each of the individuals. An α of 0.5 was consistently used, assuming a 50/50 ratio. The maximum likelihood in the mixed-genotype case was divided by the maximum likelihood in the singlet case to obtain a likelihood ratio. If this ratio was less than 1/t for some number t, the cell was assigned to be a singlet of the sample corresponding to the maximum singlet likelihood. If the ratio was greater than t, the cell was assigned to be a doublet. When the ratio was in between 1/t and t, the cell was called inconclusive: no confident call could be made from which sample(s) the cell originated. The decision boundary factor t was fixed at 2. In theory, if there are n

samples in a lane, (n – 1)/n doublets can be identified using the Demuxlet algorithm, because doublets from the same individual (1/n) cannot be identified. Further details of the algorithm can be found in Kang et al.33

Using the Demuxlet algorithm, we could confidently assign the majority (99.8%) of cells to one of the individual donors (singlets) or to two different donors (doublets) (Suppl. Fig. 1a, Suppl. Table 5). Remarkably, in two out of eight sample pools, no cells were assigned to one of the six donors within the pool. Moreover, the detected doublet rate in those sample pools was abnormally high (17.5% and 21.1%, while 3-4% was expected) (Suppl. Table 5). This is most probably due to a sample mix-up in the lab which resulted in an artificially high doublet rate. Since the genotypes of these two mixed-up samples were not available, those samples were excluded from the analysis (marked as “doublet”).

Two additional tests were performed to confirm the correct assignment of cells using Demuxlet. First, we determined what would happen if the cells did not match with their genotypes by taking six random genotypes not present in the sample pool itself. This resulted in 0.02% of the cells being a singlet, 0.03% being inconclusive and 99.95% being a doublet. Second, the number of reads mapping to the Y-chromosome was determined for the singlets of each donor. Cells belonging to a female donor showed (almost) no Y-reads (mismapping reads34 may explain the few sporadic Y-reads), whereas the majority of cells from male donors did (Suppl. Fig. 1b). So, the correct gender for each of the donors could be confirmed by looking at the number of Y-reads. These tests indicated that the Demuxlet method is correctly assigning cells to their respective donor and is suitable for detecting sample swaps.

Cell type classification

Version 1.4 of the R package Seurat16 was used to determine the cell types using the raw UMI counts from CellRanger. First, all genes that were not detected in ≥3 cells were removed. Cells in which >5% of the UMIs mapped to the mitochondrial-encoded genes were discarded as this can be a marker of bad quality cells; broken cells will leak cytoplasmic RNA, while the mitochondrial RNA content is retained inside the mitochondria.35 Also, cells expressing >3,500 genes were considered outliers and discarded (Suppl. Fig. 1c, Suppl. Table 6). Finally, all cells that were marked as doublet or inconclusive by the Demuxlet method were discarded. Supplementary figure 1d shows a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot36 in which all cells failing the above QCs are visualized. Library size normalization was performed on the UMI-collapsed gene expression for each barcode by scaling the total number of transcripts per cell to 10,000. The data was then log-2 transformed. In total, 25,291 cells and 19,723 genes (average of 1147 detected genes/cell) (see Data availability) were used in the cell type determination.

Linear regression was used to regress out the total number of UMIs and the fraction of mitochondrial transcript content per cell. The variable genes were identified using Seurat’s MeanVarPlot function which places all genes in 20 bins based on their average expression (the mean of non-zero values) and calculates the dispersion (standard deviation of all values) within each bin. Standard parameters were used except the bottom gene expression cut-off (x.low.cutoff) was set to 0 and the bottom dispersion cut-off (y.cutoff) was set to 1.0,

resulting in the identification of 1,090 genes. These 1,090 variable genes were used in the principle component analysis (PCA). The first 16 principal components were used for cell clustering using Seurat's FindCluster function (default parameters, resolution 1.2) and a t-SNE plot was used to visualize this. Based on known marker genes and differentially expressed genes per cluster (found using Seurat's FindMarkers function), we could assign 11 cell types to the clusters, including some smaller subcell types (Suppl. Fig. 2a, 2b, Suppl. Table 7). The smallest cluster we could detect consisted of plasma cells, making up 0.3% of the total PBMC population.

eQTL analysis

To find the association between genotype and expression per cell type, genome-wide *cis*-eQTL analysis for 18,264 genes (only autosomal genes, gene expressed in at least 3 cells within the total dataset and in at least 1 cell within the cell type queried, within 100 kb distance of the SNP and the gene midpoint, MAF>0.1, call rate >0.95, a Hardy-Weinberg equilibrium P value of >0.001) was performed using our previously described eQTL pipeline, version 1.2.4F (Suppl. Table 2, see Data availability).11 To assure sufficient power, cell types were merged to a more general classification: $CD4^+$ T cells, $CD8^+$ T cells, NK cells ($CD56^{dim}$ $CD16^+$ and $CD56^{bright}$ $CD16^{+/-}$), monocytes ($CD14^{bright}$ $CD16^-$ classical, cMonocyte, and $CD14^{dim}$ $CD16^+$ non-classical, ncMonocyte), B cells and DCs ($CD1C^+$ myeloid, mDC and plasmacytoid, pDC). The mean expression per gene per cell type per donor was calculated on the normalized (Z-score transformed) expression and used as input for the eQTL analysis. eQTLs were mapped using Spearman's rank correlation coefficient on imputed genotype dosages. eQTLs were considered significant at a gene-level FDR of 0.05. To control the FDR at 0.05 we used the permutation method described previously by us.2 Here we permute the link between the genotypes and expression data and create an overall null distribution using all genes. We performed in total 10 permutations and use for each gene the total null distribution of all genes to determine a gene-level FDR: during FDR estimation only the most significant SNP per gene is used, both for the real analysis and for each of the permutations.

Concordance and detection

Concordance with previously found independent top eQTLs from a whole blood DeepSAGE (3'-end transcriptomics)15 and RNA-seq study11 were computed. For this, the mean expression per gene per individual of all cells was calculated and the *cis*-eQTL mapping was confined to the independent top eQTLs found in the DeepSAGE15 or RNA-seq study11. Subsequently, detection of the same SNP-gene combination and concordance (with same allelic direction) were assessed between the significant top effects (Suppl. Table 1). Vice-versa, we also determined how many of the 379 top eQTLs in our scRNA-seq dataset could be detected and with which allelic direction within the whole blood RNA-seq study11. Similarly, we assessed detection rate and concordance with two studies containing RNA-seq data of purified cell types: Kasela et al. performed eQTL analysis on purified $CD4^+$ and $CD8^+$ T cells6, whereas the data from the Blueprint consortium contains purified $CD14^+$ monocytes and naïve $CD4^+$ T cells17 (Suppl. Table 2, 3). Moreover, for the eQTLs that were specifically detected in the cMonocytes and not the ncMonocytes (Fig. 2d), detection rate

and concordance were determined using the RNA-seq data of the purified CD14$^+$ monocytes from the Blueprint consortium17.

Single-cell gene expression imputation

To overcome the zero-inflated expression, the computational method MAGIC22 was used to impute practically all values of genes with at least some expression. MAGIC imputation (using the following parameters: 20 PCs, t=4, k=9, ka=3, ε=1) was performed per donor separately and only in the CD4$^+$ T cells (see Data availability). The effect of MAGIC imputation was validated by comparing the co-expression of typical cell type-specific marker genes (Suppl. Fig. 3).

Co-expression QTL analysis

For every individual, a Spearman's rank correlation coefficient was calculated between the expression of the *cis*-eQTL gene and all other genes. Given the large zero-inflation of scRNA-seq data, we only tested those 7,975 genes that showed variance in expression for each of the 45 samples. As a consequence we could study 102 eQTL genes out of the 145 unique genes that showed a significant *cis*-eQTL effect in CD4$^+$ T-cells. For each of these combinations, a weighted linear model was used ($co\text{-}expression \sim genotype$, where weight is $\sqrt{cellCount}$), in which the explained variable is a Spearman correlation coefficient that describes the co-expression between the two genes and the genotype is the predictor and the weights are the square root of the number of CD4$^+$ T cells within the given sample (Suppl. Fig. 5).

In order to determine for how many *cis*-eQTL genes we had identified a significant co-expression QTL we performed 100 permutations (see Data availability). For the real analysis we denoted for each of the tested 102 eQTL genes what was the most significant co-expression QTL P-Value (Suppl. Table 4). For each permutation we shuffled the genotype identifiers and reran the above analysis and also determined for each of the 102 eQTL genes what was the most significant co-expression QTL P-Value (see Data availability). This subsequently enabled us to calculate an eQTL-gene level FDR2 (using exactly the same multiple testing correction procedures as we employ for the detection of *cis-eQTLs*, see paragraph "eQTL analysis"). An eQTL gene-level FDR of 0.05 was considered significant, i.e. the p-value threshold of the most significant co-expression QTL p-values at which 5% of the co-expression QTLs are significant in the permuted compared to the real data.

All significant co-expression QTLs were discovered using non-imputed gene expression data. We then assessed whether these co-expression QTLs were also significant when using the MAGIC-imputed gene expression data. Subsequently, we tested whether these co-expression QTLs replicated using a large whole blood bulk RNA-seq dataset11 (Suppl. Table 4). We finally attempted to falsify the observed co-expression QTL for rs7297175 on the co-expression between *RPS26* and *RPL21*, by checking the following potential confounders:

- Potential sequence homology: no evidence was found for sequence homology between *RPS26* and *RPL21.*

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

- Genotype-dependent mapping problems of RNA sequence reads: no evidence was found that the *RPS26 cis*-eQTL SNP rs7297175 has any SNP proxies ($r^2 > 0.8$) that are coding and that map within *RPS26*. As such this suggests that potential genotype-dependent mapping biases of sequence-reads are unlikely.
- Multi-mapping of RNA sequence reads: no differences were found between individuals with regards to the amount of sequence reads that were discarded due to multi-mapping of sequence reads to *RPS26*.
- Unexpected *trans*-eQTL on *RPL21*: no evidence was found that the *RPS26 cis*-eQTL SNP rs7297175 is affecting the expression of *RPL21* in *trans*.
- Genotype-dependent subcell-type composition effects: the *RPS26-RPL21* co-expression QTL is unlikely the result of a subcell-type within the $CD4^+$ T cell population, as this co-expression QTL effect is also significant within $CD8^+$ T cells, within monocytes and within NK cells (Suppl. Figure 4).

Data availability

Raw gene expression counts, MAGIC imputed $CD4^+$ T cell gene expression, and eQTL and co-expression QTL summary statistics can be found under “Supplementary Data” at the website accompanying this paper (https://molgenis58.target.rug.nl/scrna-seq/).

Processed (deanonimyzed) single-cell RNA-seq data, including a text file that links each cell barcode to its respective donor, has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002560. Gene expression and genotype data can be obtained and requested by filling in a single and short web form at https://molgenis58.target.rug.nl/scrna-seq/. This form is subsequently reviewed by a single Data Access Committee, who will be able to approve access to both the raw gene expression and genotype data within 5 working days (during the holiday season there might be a slight delay). Once the proposed research is approved, access to the relevant gene expression or genotyped data will be free of charge. Access to the genotype and gene expression data is facilitated via the Lifelines workspace and the EGA, respectively. Sample metadata (age, gender, processing batch) is presented in Suppl. Table 8.

Code availability

The original R code for Seurat16 (https://github.com/satijalab/seurat), Demuxlet33 (https://github.com/statgen/demuxlet), MAGIC22 (https://github.com/KrishnaswamyLab/magic) and our in-house eQTL pipeline2 (https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline) can be found at Github. All custom-made code is made available via GitHub (https://github.com/molgenis/scRNA-seq).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are very grateful to all the volunteers who participated in this study. Moreover, we thank J. Dekens for arranging informed consent and contact with LifeLines. We thank A. Maatman and M. Platteel for their assistance in the lab. M.S and L.F. are supported by grants from the Dutch Research Council (ZonMW-VIDI 917.164.455 to M.S. and ZonMW-VIDI 917.14.374 to L.F.) and L.F. is supported by an ERC Starting Grant, grant agreement 637640 (ImmRisk). The Biobank-Based Integrative Omics Studies (BIOS) Consortium is funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007).

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012; 90:7–24. [PubMed: 22243964]
2. Westra HJ, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]
3. Brown CD, Mangravite LM, Engelhardt BE. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet. 2013; 9:e1003649. [PubMed: 23935528]
4. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012; 44:502–510. [PubMed: 22446964]
5. Fu J, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 2012; 8:e1002431. [PubMed: 22275870]
6. Kasela S, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. PLOS Genetics. 2017; 13:e1006643. [PubMed: 28248954]
7. Naranbhai V, et al. Genomic modulators of gene expression in human neutrophils. Nat Commun. 2015; 6:7545. [PubMed: 26151758]
8. Ishigaki K, et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. Nat Genet. 2017
9. Westra H, et al. Cell Specific eQTL Analysis without Sorting Cells. PLOS Genetics. 2015; 11:e1005223. [PubMed: 25955312]
10. Venet D, Pecasse F, Maenhaut C, Bersini H. Separation of samples into their constituents using gene expression data. Bioinformatics. 2001; 17(Suppl 1):S279–87. [PubMed: 11473019]
11. Zhernakova DV, et al. Identification of context-dependent expression quantitative trait loci in whole blood. Nat Genet. 2017; 49:139–145. [PubMed: 27918533]
12. Villani AC, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017; 356doi: 10.1126/science.aah4573
13. Wills QF, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. Nat Biotechnol. 2013; 31:748–752. [PubMed: 23873083]
14. Tigchelaar EF, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open. 2015; 5 e006772-2014-006772.
15. Zhernakova DV, et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genet. 2013; 9:e1003594. [PubMed: 23818875]
16. Macosko EZ, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015; 161:1202–1214. [PubMed: 26000488]
17. Chen L, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016; 167:1398–1414.e24. [PubMed: 27863251]
18. Flutre T, Wen X, Pritchard J, Stephens M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLOS Genetics. 2013; 9:e1003486. [PubMed: 23671422]
19. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet. 2013; 9:e1003491. [PubMed: 23785294]

20. Duong D, et al. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. Bioinformatics. 2017; 33:i67–i74. [PubMed: 28881962]
21. Knowles DA, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nat Methods. 2017; 14:699–702. [PubMed: 28530654]
22. van Dijk D, et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv. 2017
23. Huang M, et al. Gene Expression Recovery For Single Cell RNA Sequencing. bioRxiv. 2017
24. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018; 9:997. [PubMed: 29520097]
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
26. Simpson EH. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society.Series B (Methodological). 1951; 13:238–241.
27. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32:381–386. [PubMed: 24658644]
28. Shalek AK, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature. 2014; 510:363. [PubMed: 24919153]
29. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017; 8:14049.
30. Loh PR, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016; 48:1443–1448. [PubMed: 27694958]
31. McCarthy S, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016; 48:1279–1283. [PubMed: 27548312]
32. Das S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016; 48:1284–1287. [PubMed: 27571263]
33. Kang HM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2017
34. Rosser ZH, Balaresque P, Jobling MA. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. Am J Hum Genet. 2009; 85:130–134. [PubMed: 19576564]
35. Ilicic T, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016; 17 29-016-0888-1.
36. van de Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008; 9:2579–2605.

Europe PMC Funders Author Manuscripts

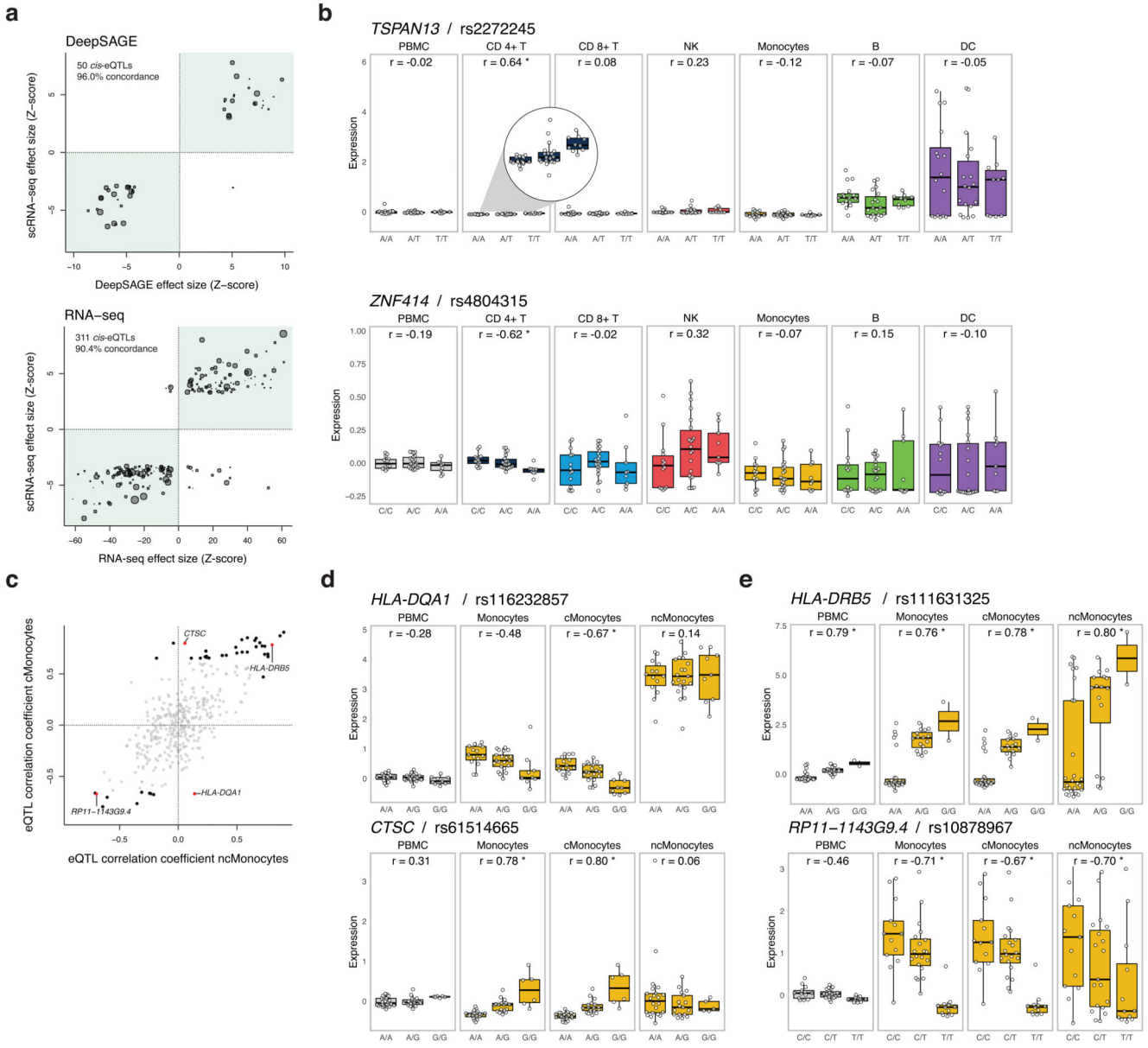


Figure 1. *Cis*-eQTL analysis in single-cell RNA-seq data.

(**a**) Effect size of the *cis*-eQTLs detected in the bulk-like PBMC scRNA-seq sample in which the analysis was confined to previously reported *cis*-eQTLs in (top) whole blood DeepSAGE or (bottom) bulk RNA-seq data. The number and percentage represent, respectively, the detected *cis*-eQTLs and their concordance (i.e. same allelic direction – green quadrants) between the bulk-like PBMC population scRNA-seq eQTLs and (top) whole blood DeepSAGE or (bottom) bulk RNA-seq data. The size of each dot represents the mean expression of the *cis*-regulated gene in the total scRNA-seq dataset. (**b**) Examples of undetectable *cis*-eQTLs in the bulk-like PBMC population caused by (top) masking of the *cis*-eQTL present in CD4$^+$ T cells but absent in DCs with comparatively high expression of the *cis*-regulated gene or (bottom) opposite allelic effects in CD4$^+$ T and NK cells. (**c**)

Europe PMC Funders Author Manuscripts

Spearman's rank correlation coefficient for the cMonocytes against the ncMonocytes of all top eQTLs that were identified in the total dataset or at least one (sub)cell cluster (see Suppl. Table 2). Significant correlations are shown in black (four red highlighted examples are shown in **d** and **e**), the non-significant in gray. (**d**) *Cis*-eQTLs specifically affecting expression in the cMonocytes, and not the ncMonocytes. (**e**) *Cis*-eQTLs significantly affecting the expression in both the cMonocytes and ncMonocytes. Each dot represents the mean expression of the eQTL gene in a donor. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range. r, Spearman's rank correlation coefficient; *FDR≤0.05.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

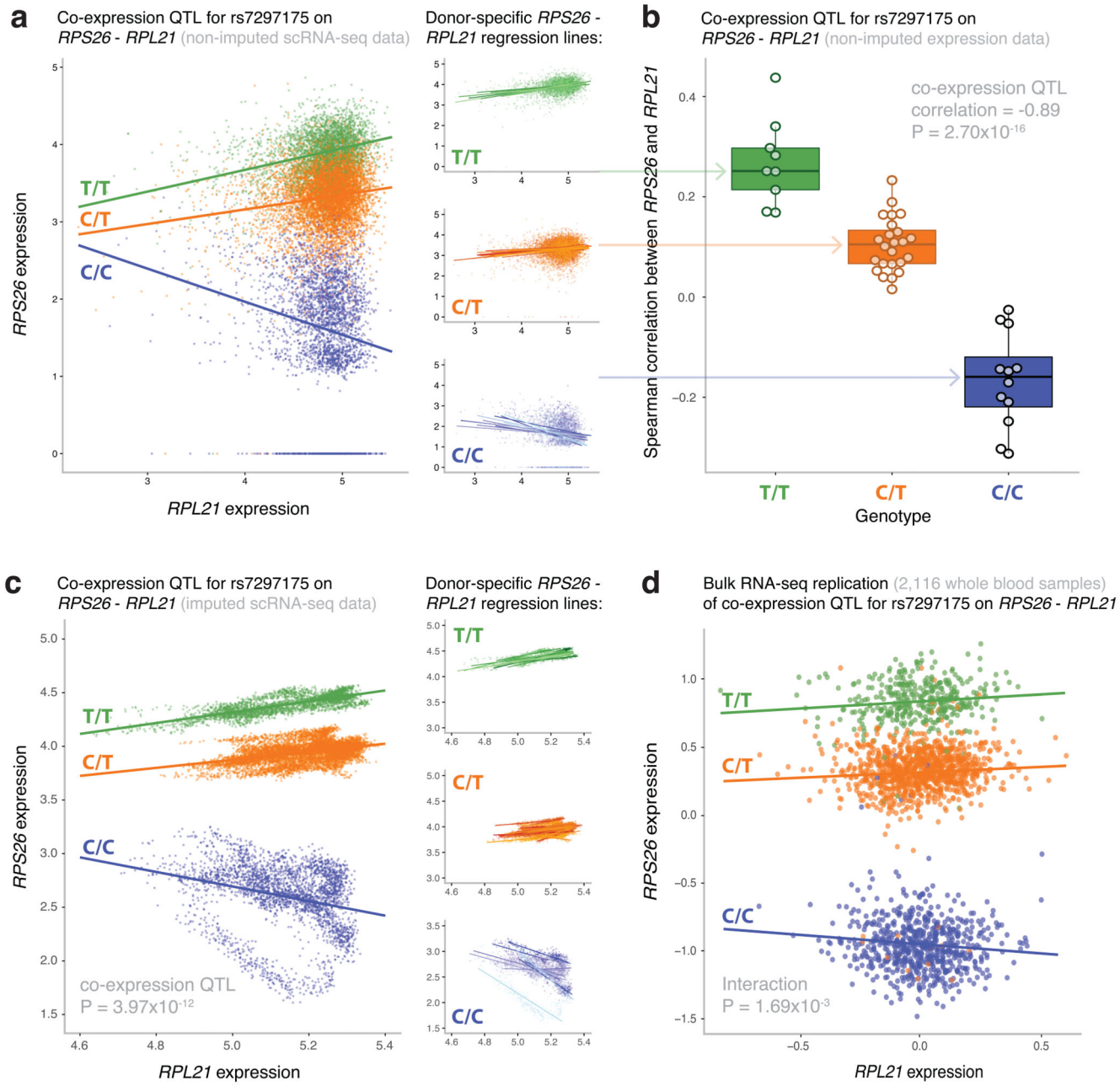


Figure 2. Most significant co-expression QTL in the CD4$^+$ T cells.

(**a**) The non-imputed expression of *RPS26* and *RPL21* of all individual CD4$^+$ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panel, respectively. Each data point represents a single cell. The nominal P-value is given for the co-expression QTL. (**b**) The Spearman's rank correlation coefficient (r) between *RPS26* and *RPL21* expression stratified by SNP rs7297175 genotype in the CD4$^+$ T cells per donor. Each data point represents a single donor. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range. The nominal P-value is given for the co-

expression QTL. (**c**) The imputed expression of *RPS26* and *RPL21* of all individual CD4$^+$ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panel, respectively. Each data point represents a single cell. (**d**) The expression of *RPS26* and *RPL21* of whole blood bulk RNA-seq samples colored by SNP rs7297175 genotype. Genotype-specific regression lines are shown. Each data point represents a single bulk RNA-seq sample. The nominal P-value is given for the interaction effect.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Cis-eQTL genes identified per cell type

Cell type	Median number of cells/donor	Unique genes with significant *cis*-eQTL effect
PBMC	507	249
CD4+ T	282	145
CD8+ T	74	21
NK	59	14
Monocyte	44	23
B	18	6
DC	11	9
Total (unique)		287

The median number of cells per donor (column 2) correlates fairly well with the number of detected *cis*-eQTL genes (column 3). In total, 379 unique top *cis*-eQTL effects, reflecting 287 unique eQTL genes, have been identified in the total dataset. Within each cell type, the number of unique *cis*-eQTL genes that we identified was equal to the number of unique, top *cis*-eQTL effects.