Supplementary Information

Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression

Urmo Võsa*[1,2#], Annique Claringbould*[1,3,4#], Harm-Jan Westra[1,3], Marc Jan Bonder[1,5], Patrick Deelen[1,3,6,7], Biao Zeng[8], Holger Kirsten[9,10], Ashis Saha[11], Roman Kreuzhuber[12,13,14], Seyhan Yazar[15], Harm Brugge[1,3], Roy Oelen[1,3], Dylan H. de Vries[1,3], Monique G.P. van der Wijst[1,3], Silva Kasela[2], Natalia Pervjakova[2], Isabel Alves[16,17], Marie-Julie Favé[16], Mawussé Agbessi[16], Mark W. Christiansen[18], Rick Jansen[19], Ilkka Seppälä[20], Lin Tong[21], Alexander Teumer[22,23], Katharina Schramm[24,25], Gibran Hemani[26], Joost Verlouw[27], Hanieh Yaghootkar[28,29,30], Reyhan Sönmez[31,32], Andrew Brown[33,34], Viktorija Kukushkina[2], Anette Kalnapenkis[2], Sina Rüeger[35], Eleonora Porcu[35], Jaanika Kronberg[2], Johannes Kettunen[36,37,38,39], Bernett Lee[40], Futao Zhang[41], Ting Qi[41], Jose Alquicira Hernandez[42], Wibowo Arindrarto[43], Frank Beutner[44], BIOS Consortium†, i2QTL Consortium†, Julia Dmitrieva[45], Mahmoud Elansary[45], Benjamin P. Fairfax[46], Michel Georges[45], Bastiaan T. Heijmans[43], Alex W. Hewitt[47,48], Mika Kähönen[49], Yungil Kim[50,11], Julian C. Knight[46], Peter Kovacs[51], Knut Krohn[52], Shuang Li[1,6], Markus Loeffler[9,10], Urko M. Marigorta[53,54,55], Hailang Mei[56], Yukihide Momozawa[45,57], Martina Müller-Nurasyid[24,25,58], Matthias Nauck[23,59], Michel G. Nivard[60], Brenda WJH. Penninx[19], Jonathan K. Pritchard[61,62], Olli T. Raitakari[63,64,65], Olaf Rotzschke[40], Eline P. Slagboom[43], Coen D.A. Stehouwer[66], Michael Stumvoll[67], Patrick Sullivan[68], Peter A.C. 't Hoen[69], Joachim Thiery[70,10], Anke Tönjes[67], Jenny van Dongen[71], Maarten van Iterson[43], Jan H. Veldink[72], Uwe Völker[73], Robert Warmerdam[1,3], Cisca Wijmenga[1], Morris Swertz[6], Anand Andiappan[40], Grant W. Montgomery[41], Samuli Ripatti[74,75,76], Markus Perola[77], Zoltan Kutalik[78], Emmanouil Dermitzakis[32,33,79], Sven Bergmann[31,32], Timothy Frayling[28], Joyce van Meurs[27], Holger Prokisch[80,81], Habibul Ahsan[21], Brandon L. Pierce[21], Terho Lehtimäki[20], Dorret I. Boomsma[71], Bruce M. Psaty[82], Sina A. Gharib[18,83], Philip Awadalla[16], Lili Milani[2], Willem H. Ouwehand[12,13,84], Kate Downes[12,13], Oliver Stegle[5,14,85], Alexis Battle[11,86], Peter M. Visscher[41], Jian Yang[41,87,88], Markus Scholz[9,10], Joseph Powell**[15,89], Greg Gibson**[53], Tõnu Esko**[2], Lude Franke**[1,3#]

* These authors contributed equally
** These authors jointly supervised
† Full author list for consortium authors appears at the end of **Supplementary Note**
# Correspondence to: Urmo Võsa (urmo.vosa@gmail.com), Annique Claringbould (anniqueclaringbould@gmail.com) and Lude Franke (lude@ludesign.nl)

1. Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
2. Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia

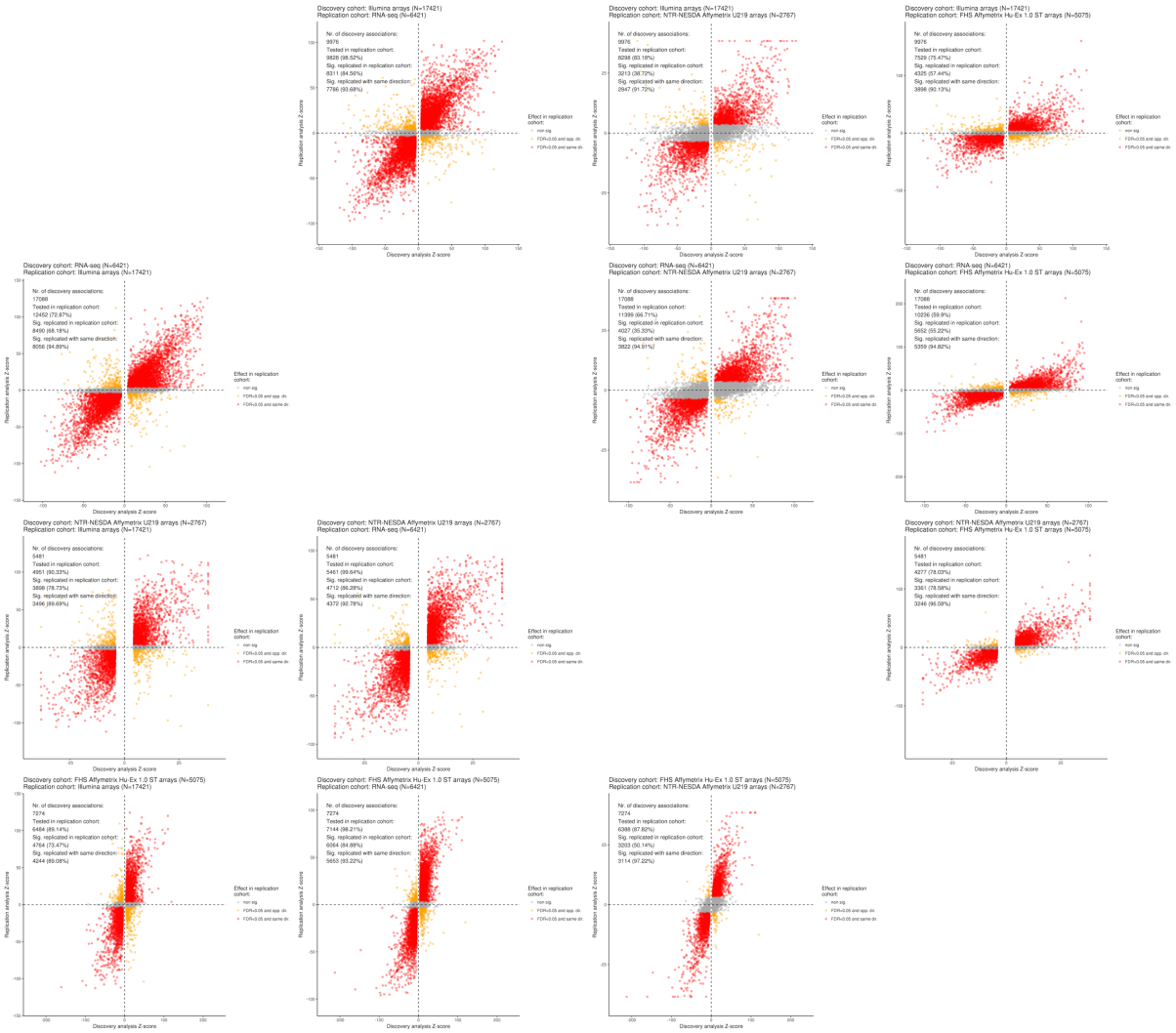
3. Oncode Institute, Amsterdam, The Netherlands
4. European Molecular Biology Laboratory, Structural & Computational Biology Unit, Heidelberg, Germany
5. European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany
6. Genomics Coordination Center, University Medical Centre Groningen, Groningen, The Netherlands
7. Department of Genetics, University Medical Centre Utrecht, Utrecht, The Netherlands
8. School of Biological Sciences, Georgia Tech, Atlanta, United States of America
9. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany
10. LIFE Research Center for Civilization Diseases, University of Leipzig, Leipzig, Germany
11. Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, United States of America
12. Department of Haematology, University of Cambridge, Cambridge, United Kingdom
13. NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom
14. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
15. Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, New South Wales, Australia
16. Computational Biology, Ontario Institute for Cancer Research, Toronto, Ontario, Canada
17. l'institut du thorax, Université de Nantes, CHU Nantes, INSERM, CNRS, Nantes, France
18. Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, United States of America
19. Department of Psychiatry, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute and Amsterdam Neuroscience, Amsterdam, The Netherlands
20. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
21. Department of Public Health Sciences, University of Chicago, Chicago, Illinois, United States of America
22. Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
23. DZHK (German Center for Cardiovascular Research), partner site Greifswald, Greifswald, Germany
24. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany
25. Department of Medicine I, University Hospital Munich, Ludwig Maximilian's University, Munich, Germany
26. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
27. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands
28. Genetics of Complex Traits, University of Exeter Medical School, Royal Devon & Exeter Hospital, Exeter, United Kingdom
29. School of Life Sciences, College of Liberal Arts and Science, University of Westminster, London, United Kingdom
30. Division of Medical Sciences, Department of Health Sciences, Luleå University of Technology, Luleå, Sweden
31. Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
32. Swiss Institute of Bioinformatics, Lausanne, Switzerland
33. Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
34. Population Health and Genomics, University of Dundee, Dundee, United Kingdom
35. Lausanne University Hospital, Lausanne, Switzerland
36. Computational Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland
37. Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland
38. Biocenter Oulu, University of Oulu, Oulu, Finland
39. Finnish Institute for Health and Welfare, Helsinki, Finland

40. Singapore Immunology Network, Agency for Science, Technology and Research, Singapore, Singapore
41. Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia
42. Garvan Institute of Medical Research, Garvan-Weizmann Centre for Cellular Genomics, Sydney, Australia
43. Leiden University Medical Center, Leiden, The Netherlands
44. Heart Center Leipzig, Universität Leipzig, Leipzig, Germany
45. Unit of Animal Genomics, WELBIO, GIGA-R & Faculty of Veterinary Medicine, University of Liege, Liège, Belgium
46. Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom
47. Menzies Institute for Medical Research, School of Medicine, University of Tasmania, Hobart, Tasmania, Australia
48. Centre for Eye Research Australia, Department of Surgery, University of Melbourne, Melbourne, Victoria, Australia
49. Department of Clinical Physiology, Tampere University Hospital and Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
50. Genetics and Genomic Science Department, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America
51. IFB Adiposity Diseases, Universität Leipzig, Leipzig, Germany
52. Interdisciplinary Center for Clinical Research, Faculty of Medicine, Universität Leipzig, Leipzig, Germany
53. School of Biological Sciences, Georgia Tech, Atlanta, Georgia, United States of America
54. Integrative Genomics Lab, CIC bioGUNE, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Derio, Bizkaia, Basque Country, Spain
55. IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
56. Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
57. Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Kanagawa, Japan
58. IBE, Faculty of Medicine, LMU Munich, Munich, Germany
59. Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany
60. Department of Biological Psychology, Faculty of Behaviour and Movement Sciences, VU, Amsterdam, The Netherlands
61. Department of Biology, Stanford University, Stanford, California, United States of America
62. Department of Genetics, Stanford University, Stanford, California, United States of America
63. Centre for Population Health Research, University of Turku and Turku University Hospital, Turku, Finland
64. Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland
65. Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland
66. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
67. Department of Medicine, Universität Leipzig, Leipzig, Germany
68. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
69. Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, The Netherlands
70. Institute for Laboratory Medicine, LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig, Leipzig, Germany
71. Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam Public Health research institute and Amsterdam Neuroscience, Amsterdam, the Netherlands

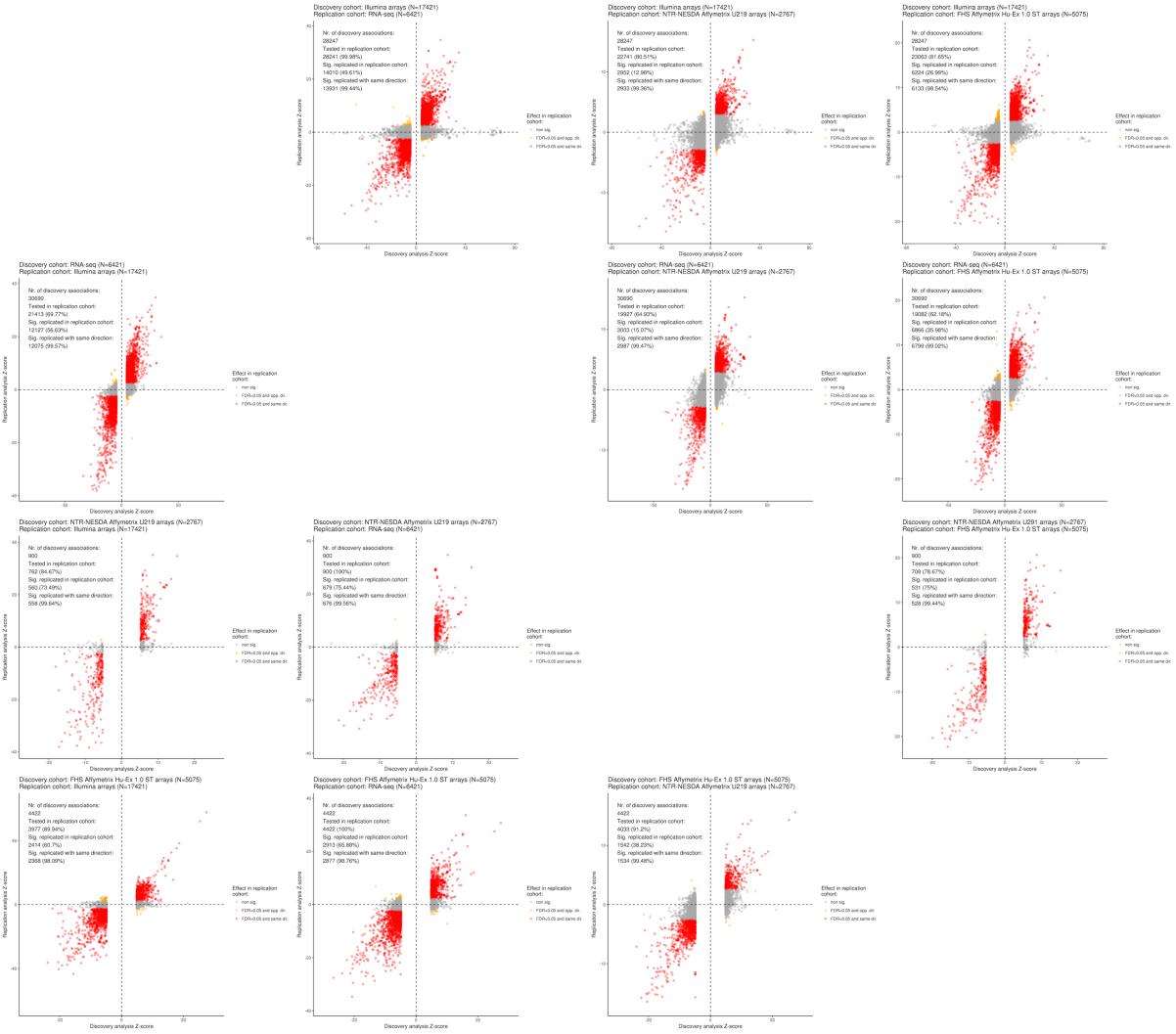
72. UMC Utrecht Brain Center, University Medical Center Utrecht, Department of Neurology, Utrecht University, Utrecht, The Netherlands
73. Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany
74. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland
75. Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland
76. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America
77. National Institute for Health and Welfare, University of Helsinki, Helsinki, Finland
78. Center for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland
79. Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland
80. Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
81. Institute of Human Genetics, Technical University Munich, Munich, Germany
82. Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, Washington, USA
83. Department of Medicine, University of Washington, Seattle, Washington, United States of America
84. Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
85. Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany
86. Departments of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, United States of America
87. School of Life Sciences, Westlake University, Hangzhou, China
88. Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China
89. UNSW Cellular genomics Futures Institute, University of New South Wales, Sydney, New South Wales, Australia

Contents

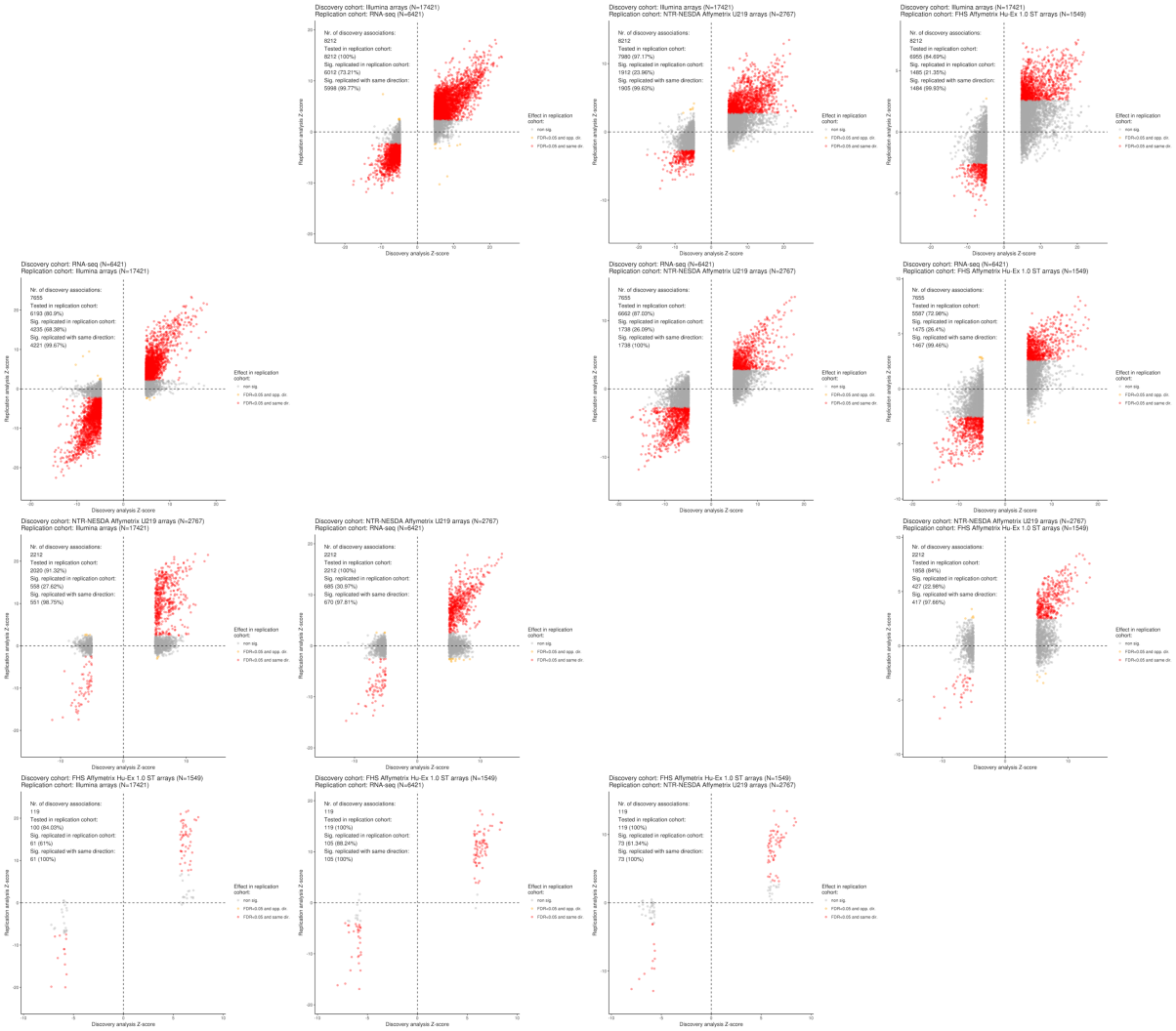
Supplementary Figures



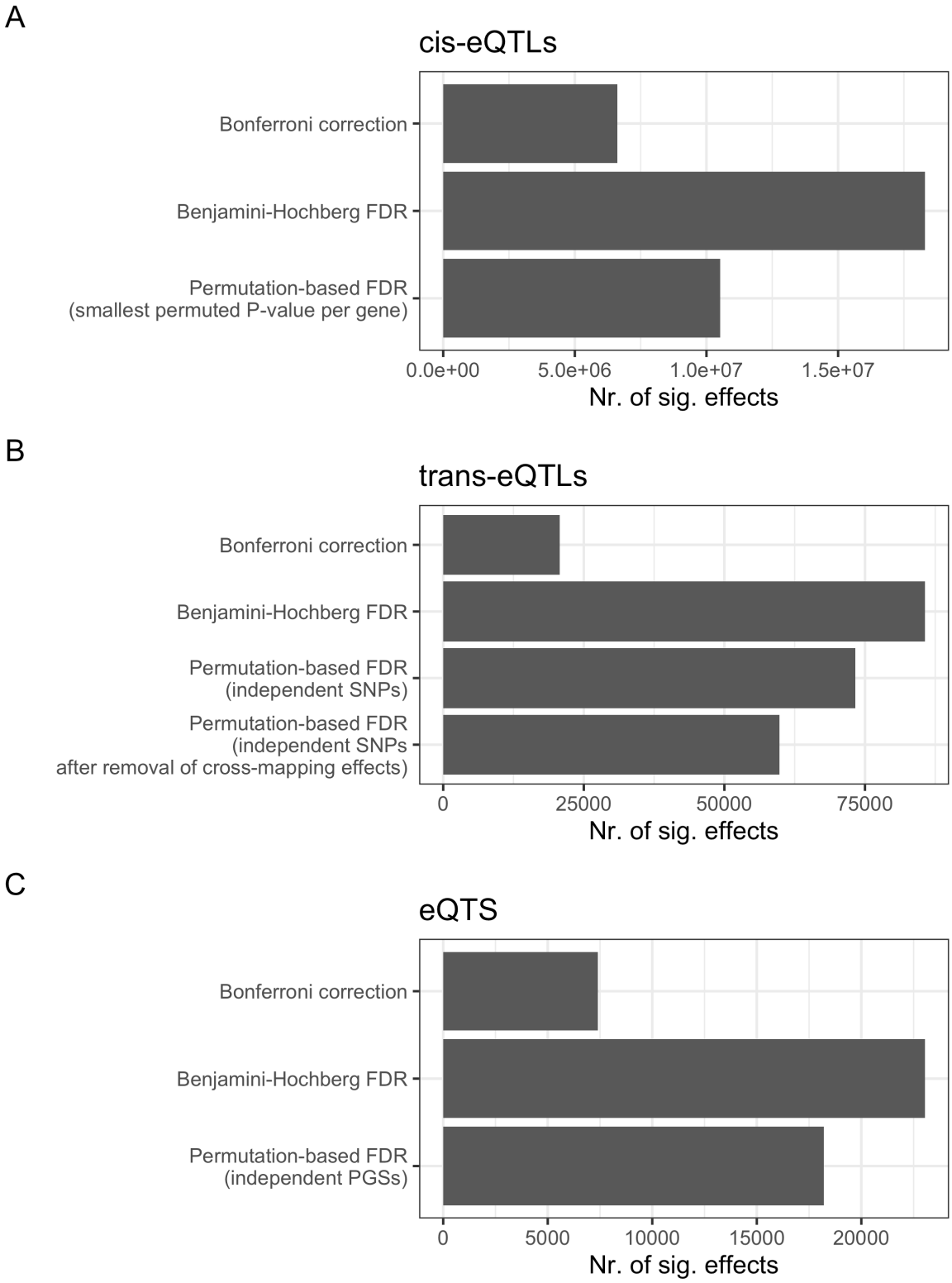
Supplementary Figure 1A. Between-platform replications for *cis*-eQTLs. In this analysis, discovery *cis*-eQTL lead SNP for each gene was compared against all effects in the replication dataset. Note that for better visualization, scales of x- and y-axis vary on each plot. Grey dots indicate eQTLs not significant in the replication dataset (FDR≥0.05), yellow dots indicate eQTLs significant in the replication dataset (FDR<0.05) but with opposite allelic effect, and red dots indicate eQTLs significant in the replication dataset with identical allelic effect direction as in the discovery dataset. *Cis*-eQTLs show high concordance between gene expression platforms.



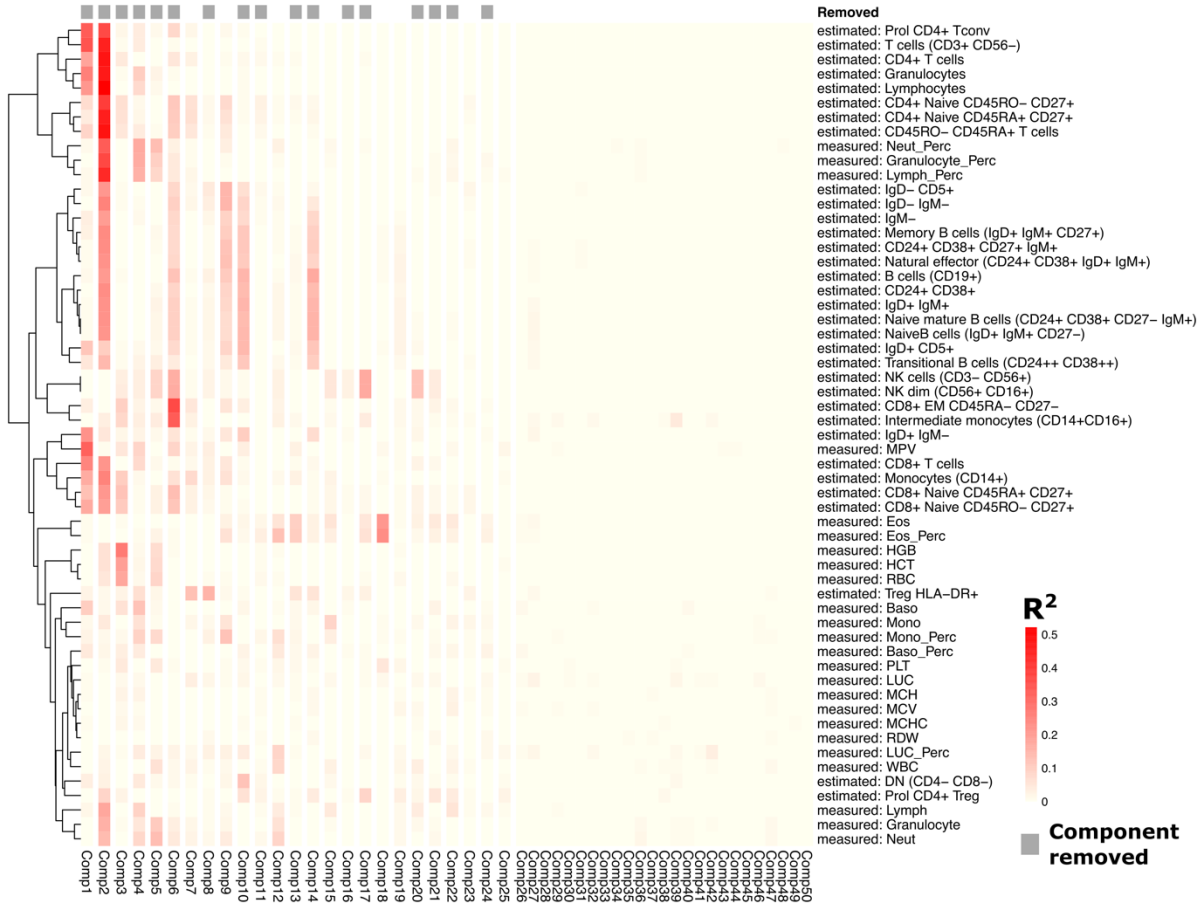
Supplementary Figure 1B. Between-platform replications for *trans*-eQTLs. Note that for better visualization, scales of x- and y-axis vary on each plot. Grey dots indicate eQTLs not significant in the replication dataset (FDR≥0.05), yellow dots indicate eQTLs significant in the replication dataset (FDR<0.05) but with opposite allelic effect, and red dots indicate eQTLs significant in the replication dataset with identical allelic effect direction as in the discovery dataset. *Trans*-eQTLs show high concordance between gene expression platforms.



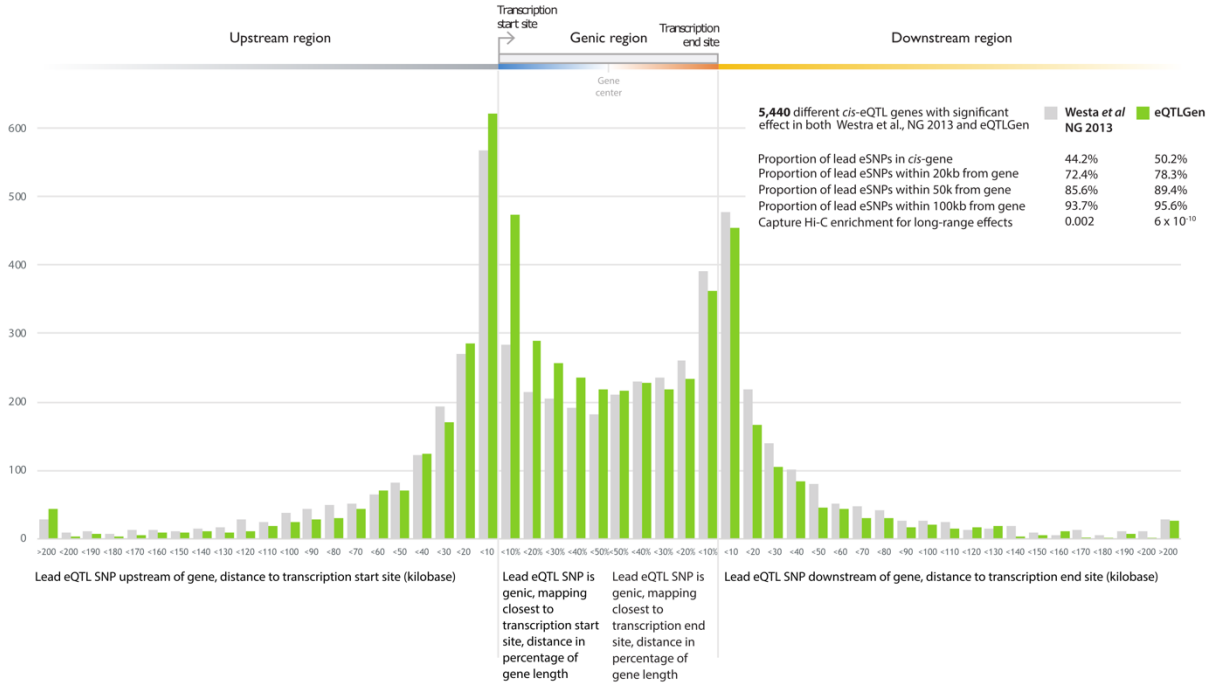
Supplementary Figure 1C. Between-platform replications for eQTSs. Note that for better visualization, scales of x- and y-axis vary on each plot. Grey dots indicate eQTSs not significant in the replication dataset (FDR≥0.05), yellow dots indicate eQTSs significant in the replication dataset (FDR<0.05) but with opposite allelic effect, and red dots indicate eQTSs significant in the replication dataset with identical allelic effect direction as in the discovery dataset. eQTSs show high concordance between gene expression platforms.



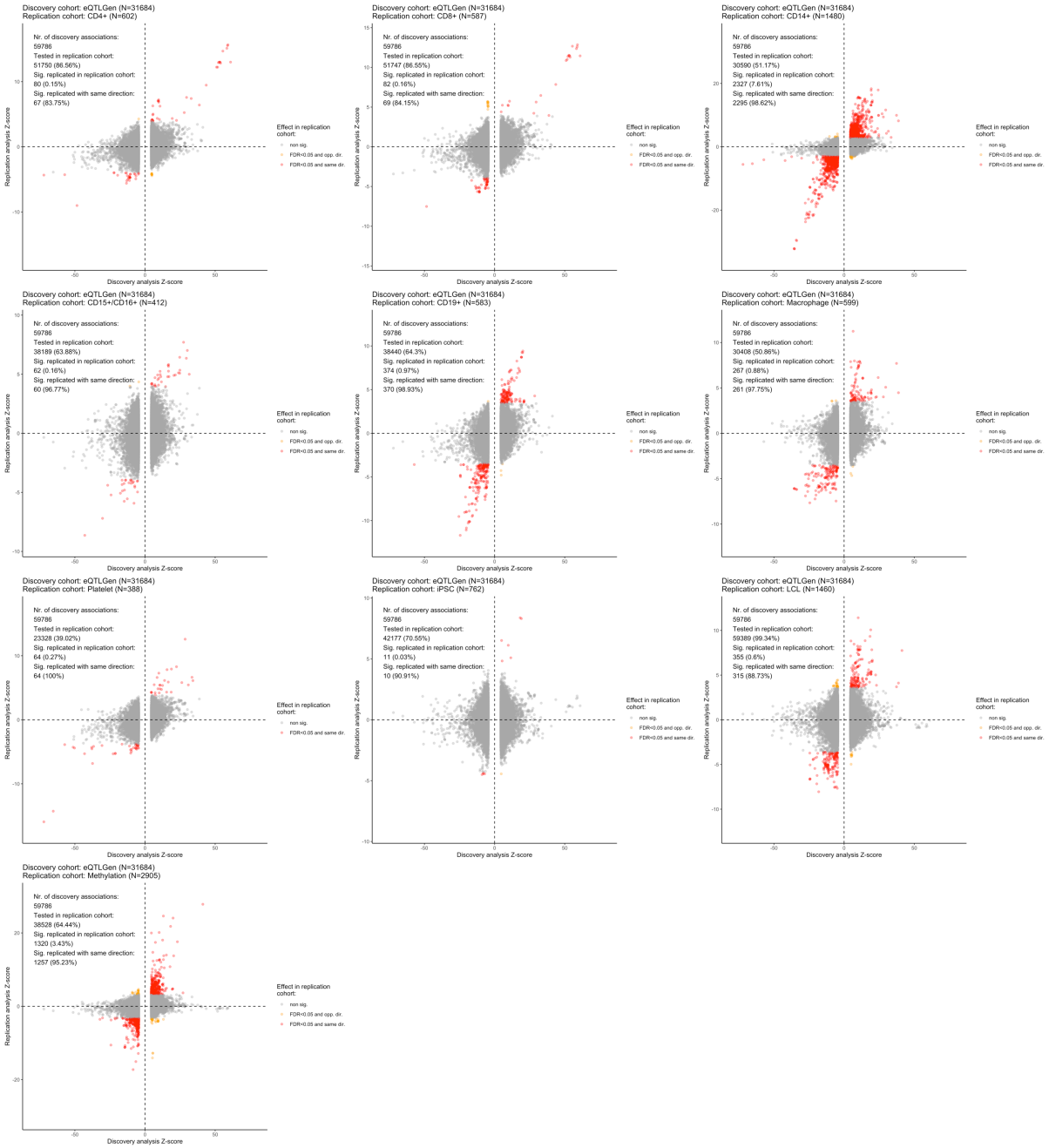
Supplementary Figure 2. Comparison of the numbers of significant *cis*-eQTL, *trans*-eQTL and eQTS effects acquired by different multiple testing methods. Bonferroni corrected P-values and Benjamini-Hochberg FDRs were calculated by p.adjust() command in R, permutation-based FDRs were calculated with our pipeline as specified in **Methods**.



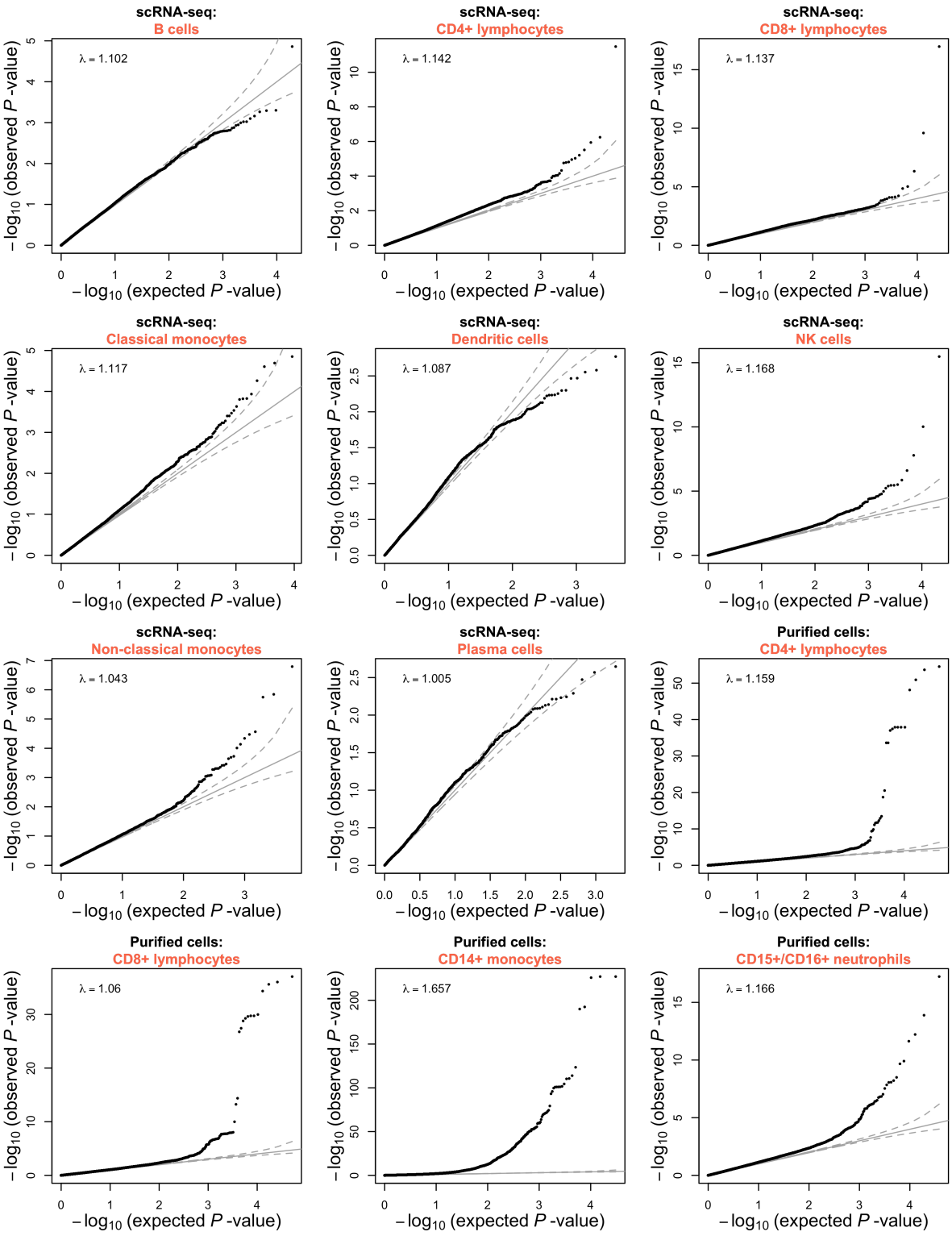
Supplementary Figure 3. Correlations between gene expression principal components and cell metrics in BIOS cohorts. Cell metrics were measured directly or estimated by Decon-cell (part of Decon2 framework). Visualised are squared Spearman correlation coefficients and no transformation was applied on cell metrics. Each correlation was calculated using the maximum number of samples available for this specific comparison (N = 446–3,831). Multiple testing threshold was determined by Benjamini-Hochberg FDR over all comparisons. Grey boxes indicate principal components which were not associated with genetic variation and regressed out prior analyses.

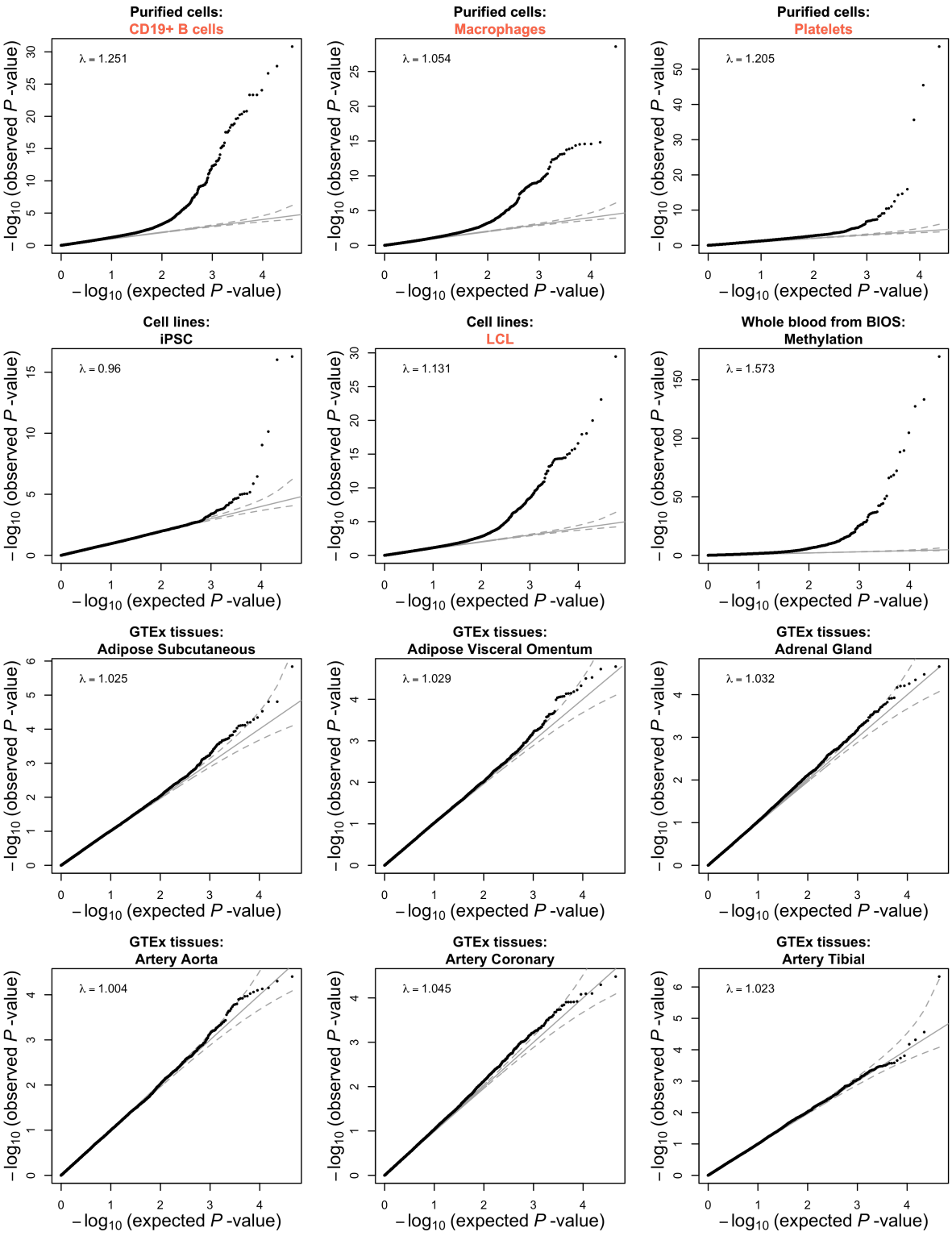


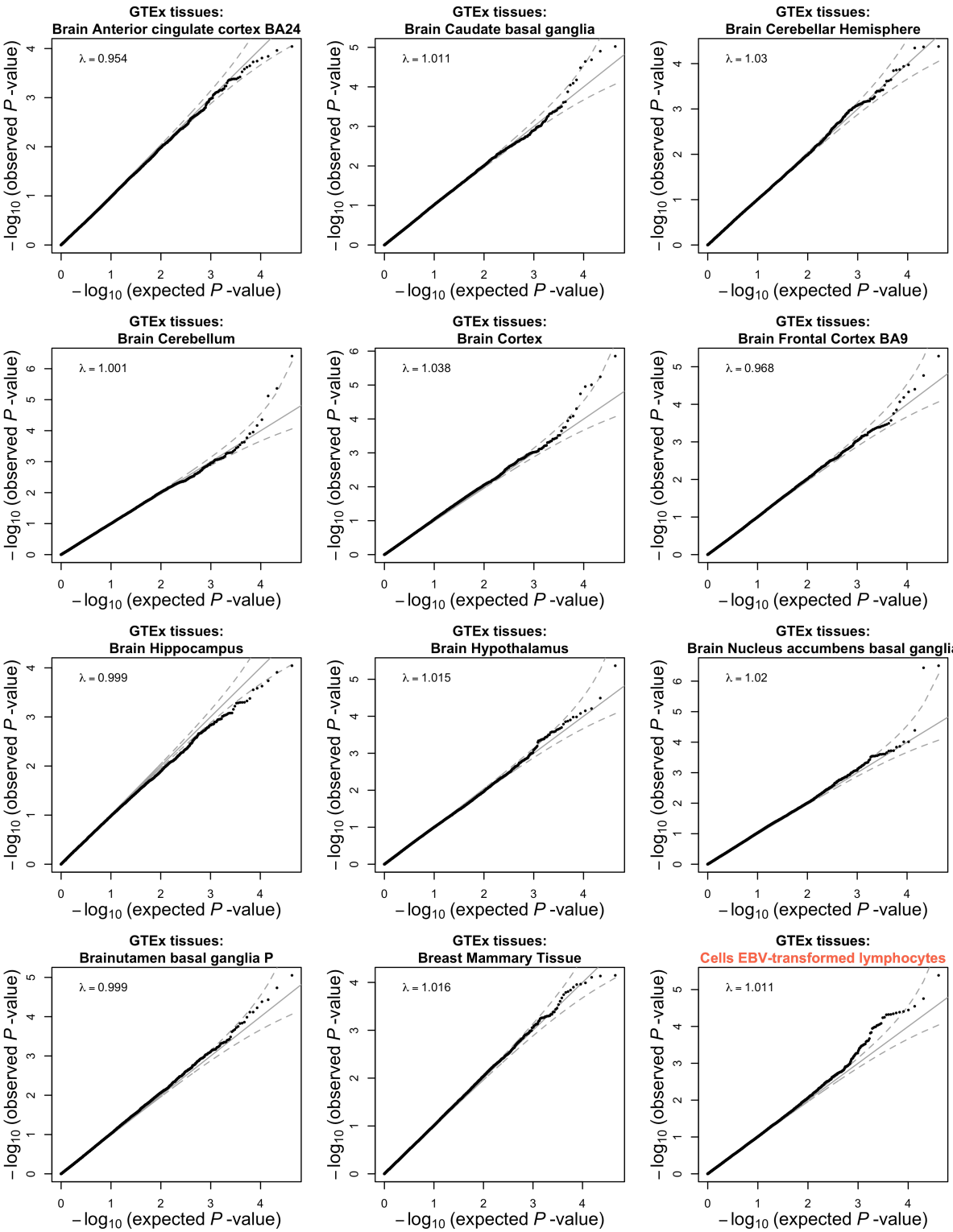
Supplementary Figure 4. Genomic positions for lead *cis*-eQTL SNPs relative to *cis*-eQTL gene positions. Compared are locations from a previous blood-based meta-analysis (Westra et al., 2013, N=5,331, in grey) and the current meta-analysis (N=31,684, in green). In the current meta-analysis results, more *cis*-eQTL lead SNPs are positioned within the gene body.

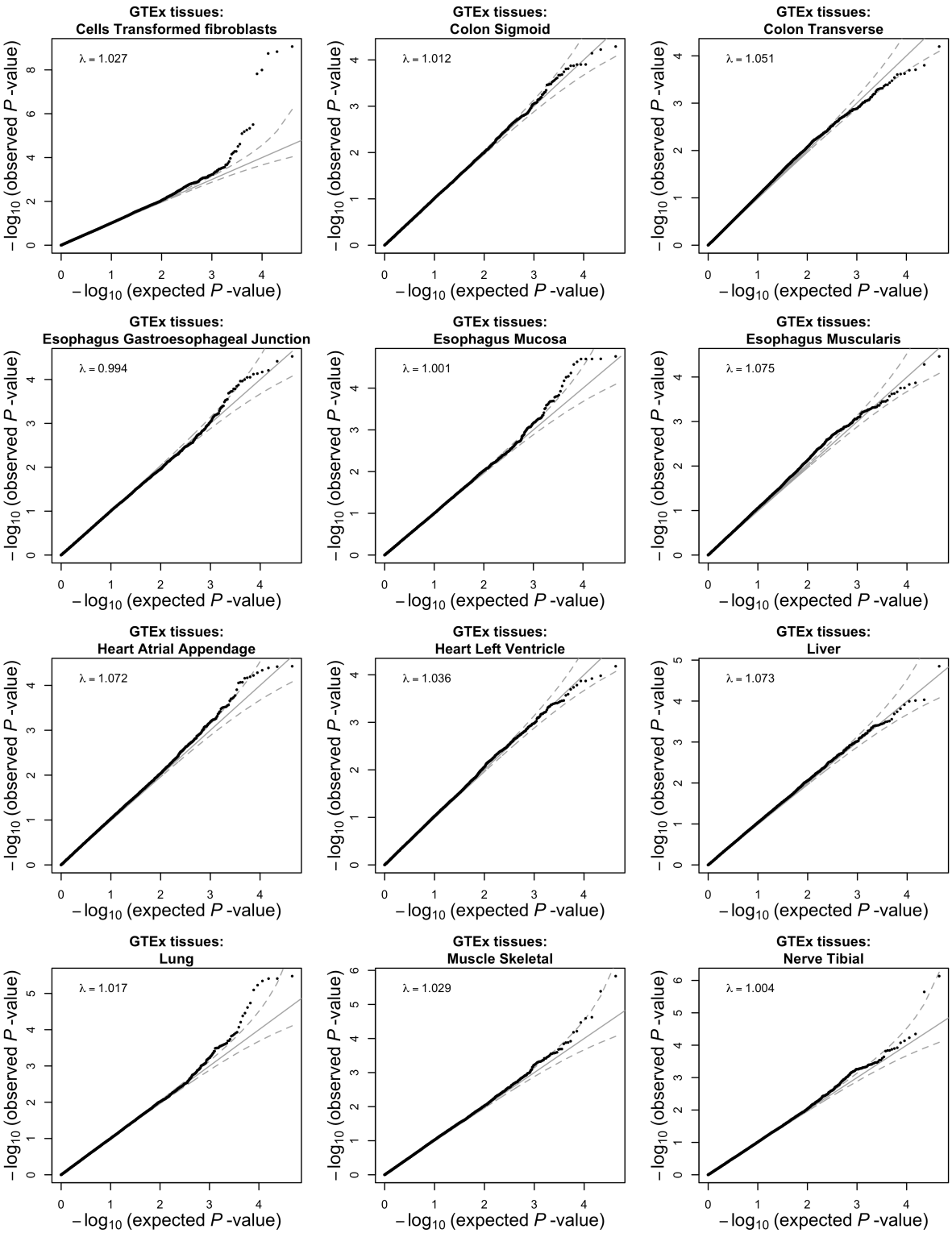


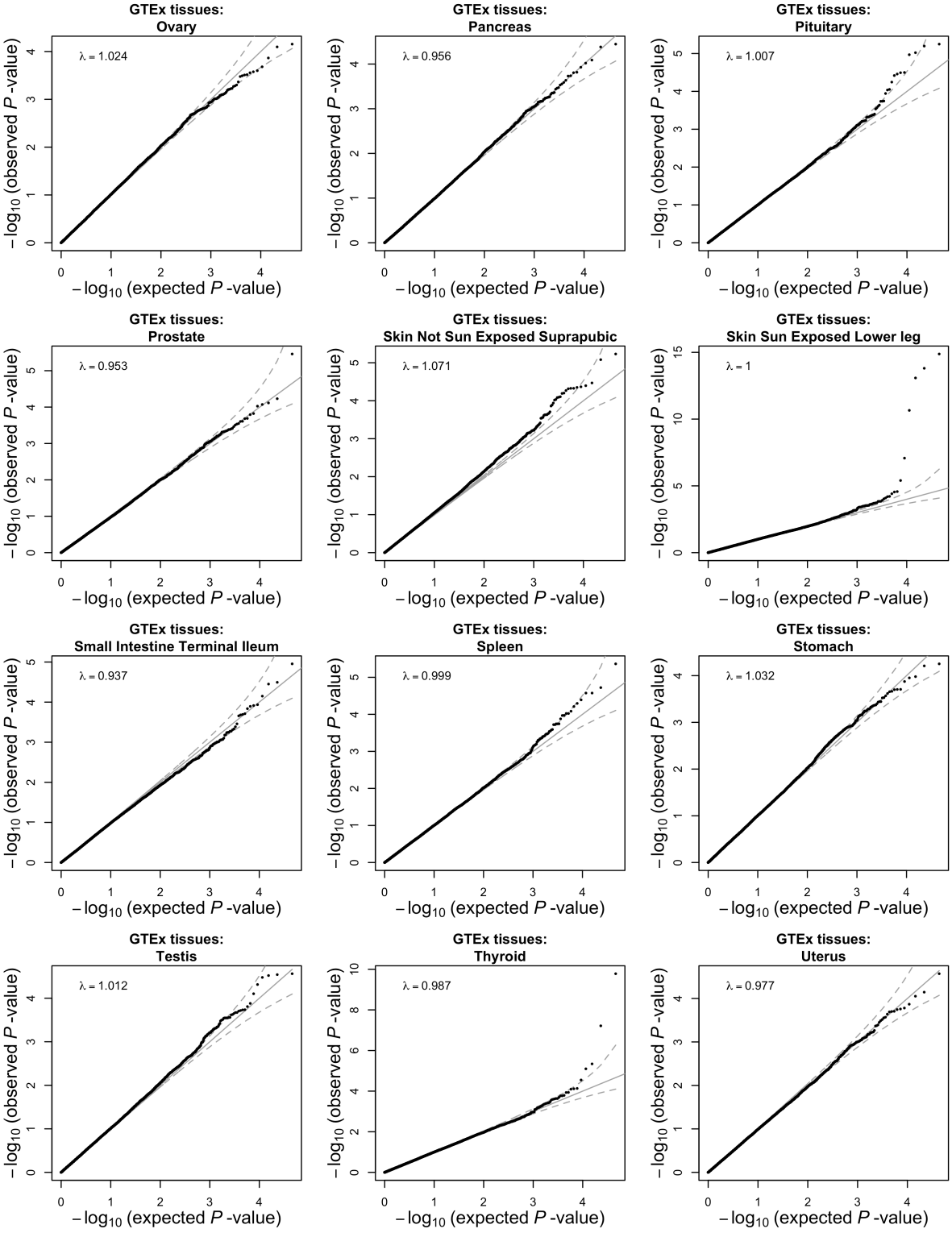
Supplementary Figure 5. Comparison of *trans*-eQTL meta-analysis Z-scores (x-axis) and replication analyses Z-scores in purified cell types, cell lines and in BIOS methylation data (y-axes). Note that for better visualization, the scale of y-axis varies on each plot. Grey dots indicate *trans*-eQTLs that were not significant (Benjamini-Hochberg FDR≥0.05) in the replication study, red dots indicate *trans*-eQTLs with significant (Benjamini-Hochberg FDR<0.05) in the replication analysis and identical allelic direction with the discovery analysis, and orange dots indicate significant (Benjamini-Hochberg FDR<0.05) effects in the replication analysis but opposite allelic direction with discovery analysis. BIOS methylation data contains samples which were part of the discovery meta-analysis and should not be considered as independent replication. Rather it indicates whether *trans* effects on gene expression are showing QTL effects on different data modality (methylation).

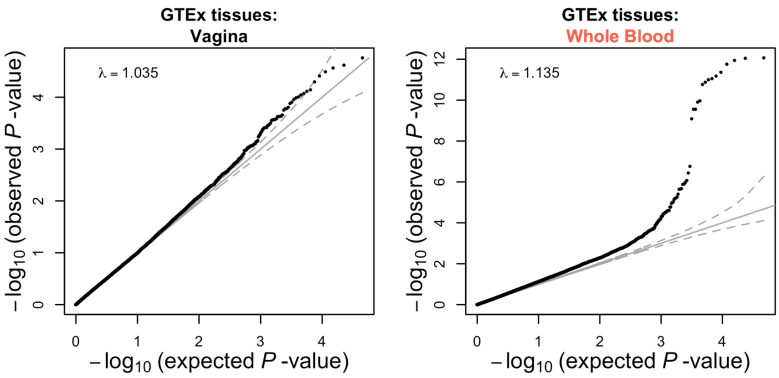

scRNA-seq:
B cells
λ = 1.102
scRNA-seq:
CD4+ lymphocytes
λ = 1.142
scRNA-seq:
CD8+ lymphocytes
λ = 1.137
scRNA-seq:
Classical monocytes
λ = 1.117
scRNA-seq:
Dendritic cells
λ = 1.087
scRNA-seq:
NK cells
λ = 1.168
scRNA-seq:
Non-classical monocytes
λ = 1.043
scRNA-seq:
Plasma cells
λ = 1.005
Purified cells:
CD4+ lymphocytes
λ = 1.159
Purified cells:
CD8+ lymphocytes
λ = 1.06
Purified cells:
CD14+ monocytes
λ = 1.657
Purified cells:
CD15+/CD16+ neutrophils
λ = 1.166
− log10 (observed P -value)
− log10 (expected P -value)


Purified cells:
CD19+ B cells
λ = 1.251
Purified cells:
Macrophages
λ = 1.054
Purified cells:
Platelets
λ = 1.205
Cell lines:
iPSC
λ = 0.96
Cell lines:
LCL
λ = 1.131
Whole blood from BIOS:
Methylation
λ = 1.573
GTEx tissues:
Adipose Subcutaneous
λ = 1.025
GTEx tissues:
Adipose Visceral Omentum
λ = 1.029
GTEx tissues:
Adrenal Gland
λ = 1.032
GTEx tissues:
Artery Aorta
λ = 1.004
GTEx tissues:
Artery Coronary
λ = 1.045
GTEx tissues:
Artery Tibial
λ = 1.023
− log10 (observed P -value)
− log10 (expected P -value)

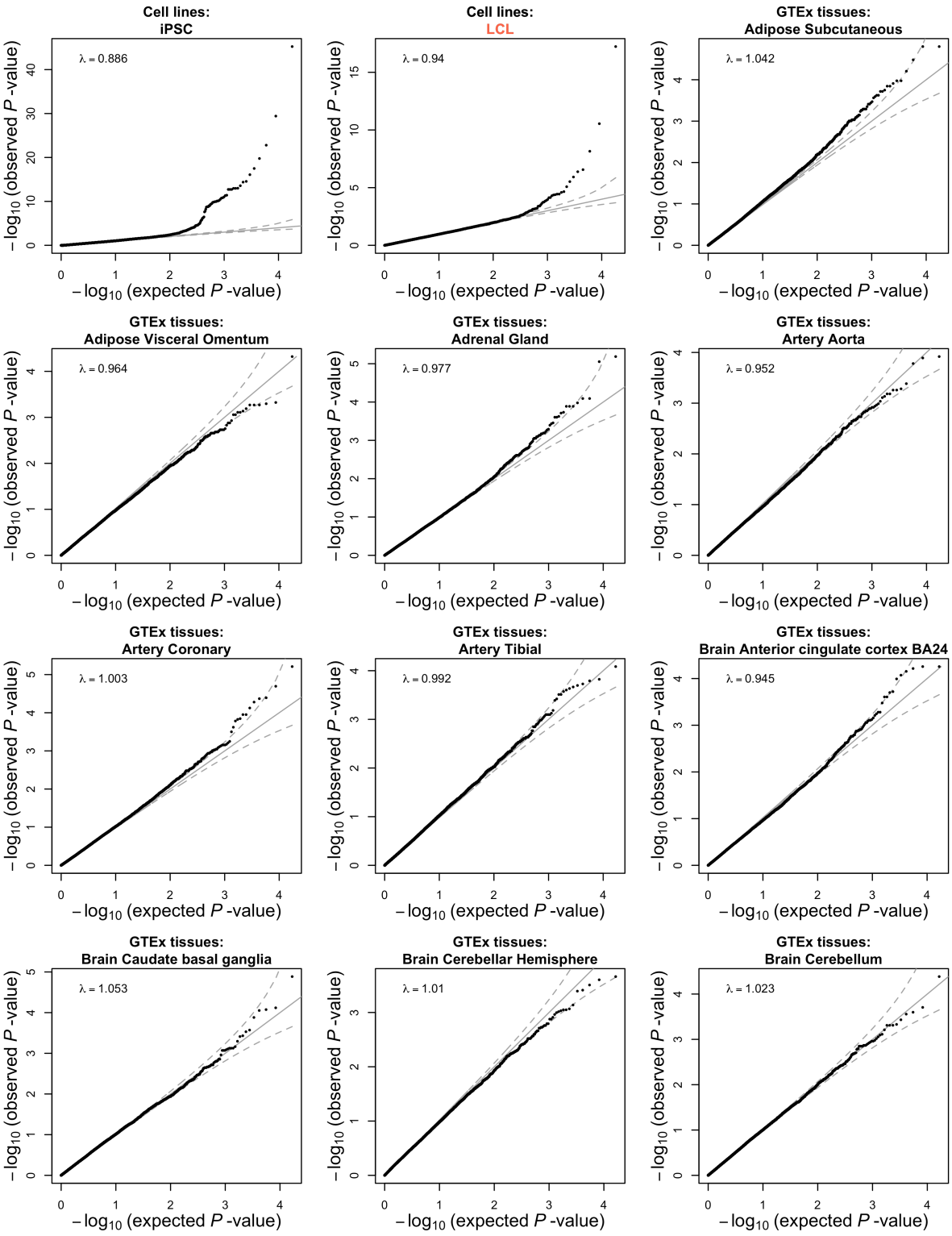

GTEx tissues:
Brain Anterior cingulate cortex BA24
λ = 0.954
GTEx tissues:
Brain Caudate basal ganglia
λ = 1.011
GTEx tissues:
Brain Cerebellar Hemisphere
λ = 1.03
GTEx tissues:
Brain Cerebellum
λ = 1.001
GTEx tissues:
Brain Cortex
λ = 1.038
GTEx tissues:
Brain Frontal Cortex BA9
λ = 0.968
GTEx tissues:
Brain Hippocampus
λ = 0.999
GTEx tissues:
Brain Hypothalamus
λ = 1.015
GTEx tissues:
Brain Nucleus accumbens basal ganglia
λ = 1.02
GTEx tissues:
Brainutamen basal ganglia P
λ = 0.999
GTEx tissues:
Breast Mammary Tissue
λ = 1.016
GTEx tissues:
Cells EBV-transformed lymphocytes
λ = 1.011
− log10 (observed P -value)
− log10 (expected P -value)

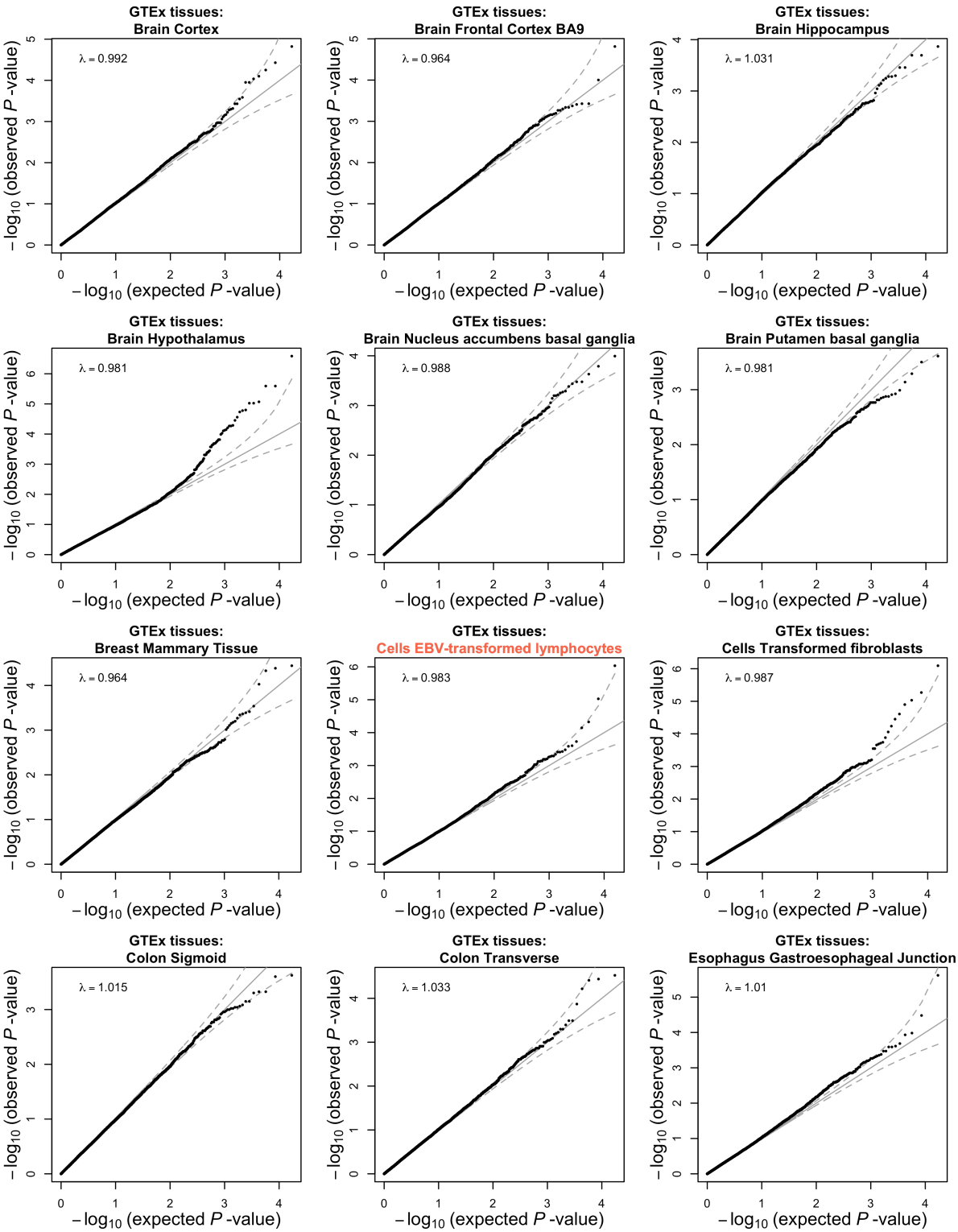

GTEx tissues:
Cells Transformed fibroblasts
λ = 1.027
GTEx tissues:
Colon Sigmoid
λ = 1.012
GTEx tissues:
Colon Transverse
λ = 1.051
GTEx tissues:
Esophagus Gastroesophageal Junction
λ = 0.994
GTEx tissues:
Esophagus Mucosa
λ = 1.001
GTEx tissues:
Esophagus Muscularis
λ = 1.075
GTEx tissues:
Heart Atrial Appendage
λ = 1.072
GTEx tissues:
Heart Left Ventricle
λ = 1.036
GTEx tissues:
Liver
λ = 1.073
GTEx tissues:
Lung
λ = 1.017
GTEx tissues:
Muscle Skeletal
λ = 1.029
GTEx tissues:
Nerve Tibial
λ = 1.004
− log10 (observed P -value)
− log10 (expected P -value)

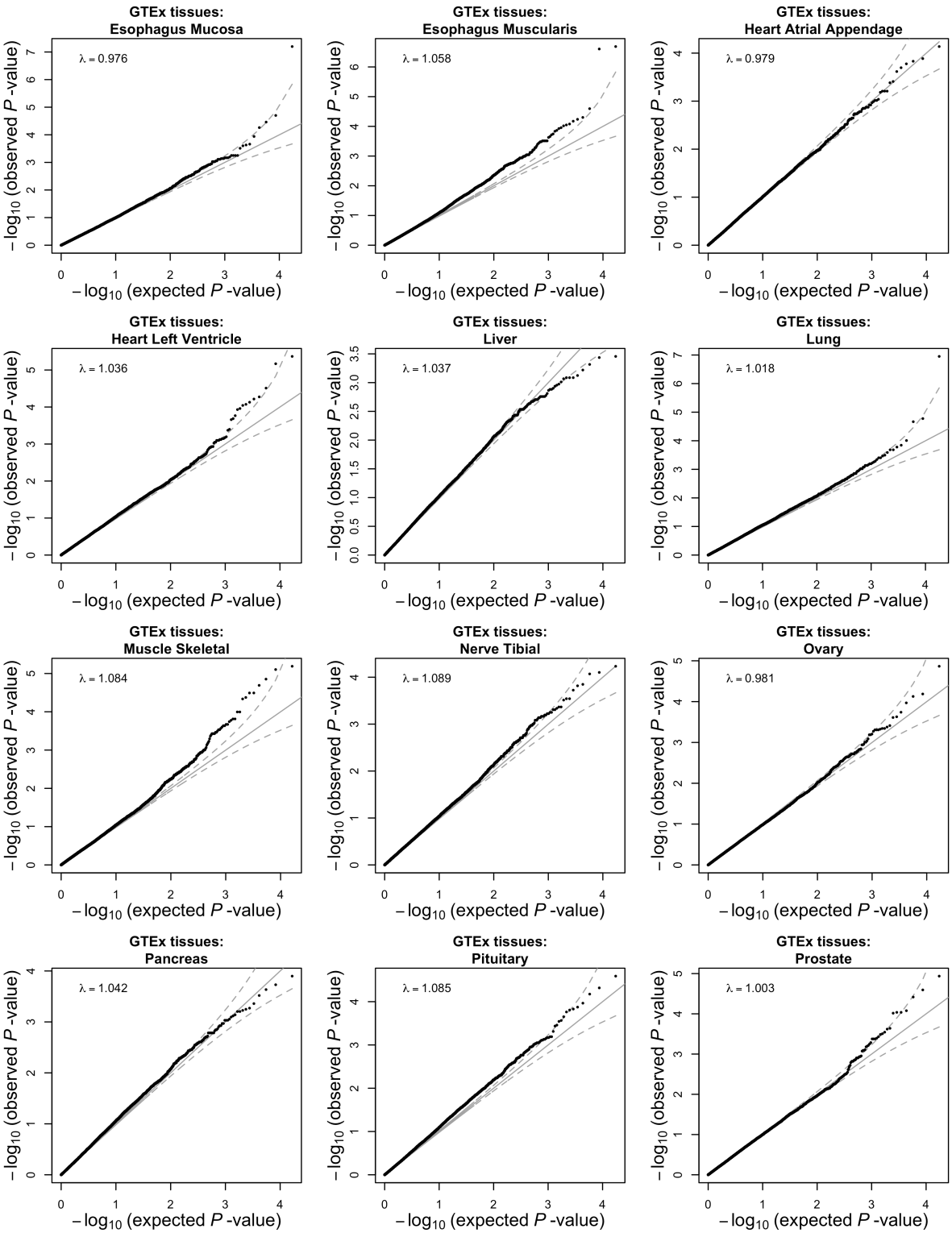

GTEx tissues:
Ovary
λ = 1.024
GTEx tissues:
Pancreas
λ = 0.956
GTEx tissues:
Pituitary
λ = 1.007
GTEx tissues:
Prostate
λ = 0.953
GTEx tissues:
Skin Not Sun Exposed Suprapubic
λ = 1.071
GTEx tissues:
Skin Sun Exposed Lower leg
λ = 1
GTEx tissues:
Small Intestine Terminal Ileum
λ = 0.937
GTEx tissues:
Spleen
λ = 0.999
GTEx tissues:
Stomach
λ = 1.032
GTEx tissues:
Testis
λ = 1.012
GTEx tissues:
Thyroid
λ = 0.987
GTEx tissues:
Uterus
λ = 0.977
− log10 (observed P -value)
− log10 (expected P -value)

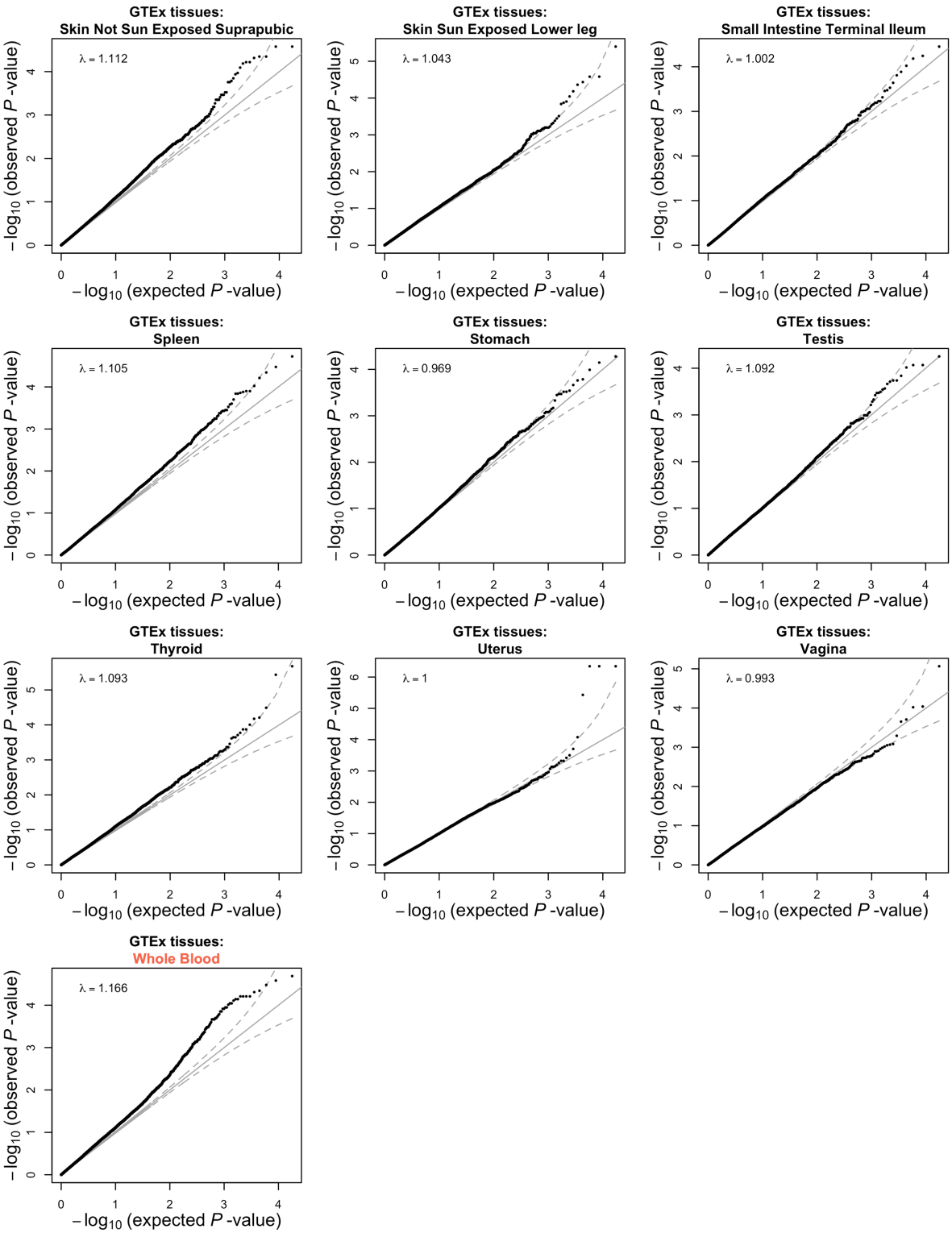


Supplementary Figure 6A. QQ-plots for investigating the inflation of *trans*-eQTL signal in different replication datasets: scRNA-seq datasets, purified cell types, cell lines, bulk tissues from GTEx v6p and BIOS methylation data. Each plot visualises two-sided -$\log_{10}$(P-values) from each replication analysis (Spearman correlation; per-dataset effects were meta-analyzed if replication included multiple datasets from the same cell- or tissue type; **Methods**, **Supplementary Methods**). Replication datasets from whole blood and from individual immune cell types are outlined with a red header. For GTEx replications, separate discovery meta-analysis was performed without GTEx, QQ-plots and lambda inflation values were calculated using *trans*-eQTL effects reaching FDR<0.05 in this separate discovery analysis. BIOS methylation data includes samples which were part of the discovery meta-analysis. The methylation replication should thus not be considered as independent replication, but rather as evidence that *trans* effects on gene expression also affect a different data layer (methylation).


Cell lines:
iPSC
λ = 0.886
Cell lines:
LCL
λ = 0.94
GTEx tissues:
Adipose Subcutaneous
λ = 1.042
GTEx tissues:
Adipose Visceral Omentum
λ = 0.964
GTEx tissues:
Adrenal Gland
λ = 0.977
GTEx tissues:
Artery Aorta
λ = 0.952
GTEx tissues:
Artery Coronary
λ = 1.003
GTEx tissues:
Artery Tibial
λ = 0.992
GTEx tissues:
Brain Anterior cingulate cortex BA24
λ = 0.945
GTEx tissues:
Brain Caudate basal ganglia
λ = 1.053
GTEx tissues:
Brain Cerebellar Hemisphere
λ = 1.01
GTEx tissues:
Brain Cerebellum
λ = 1.023
− log10 (observed P -value)
− log10 (expected P -value)

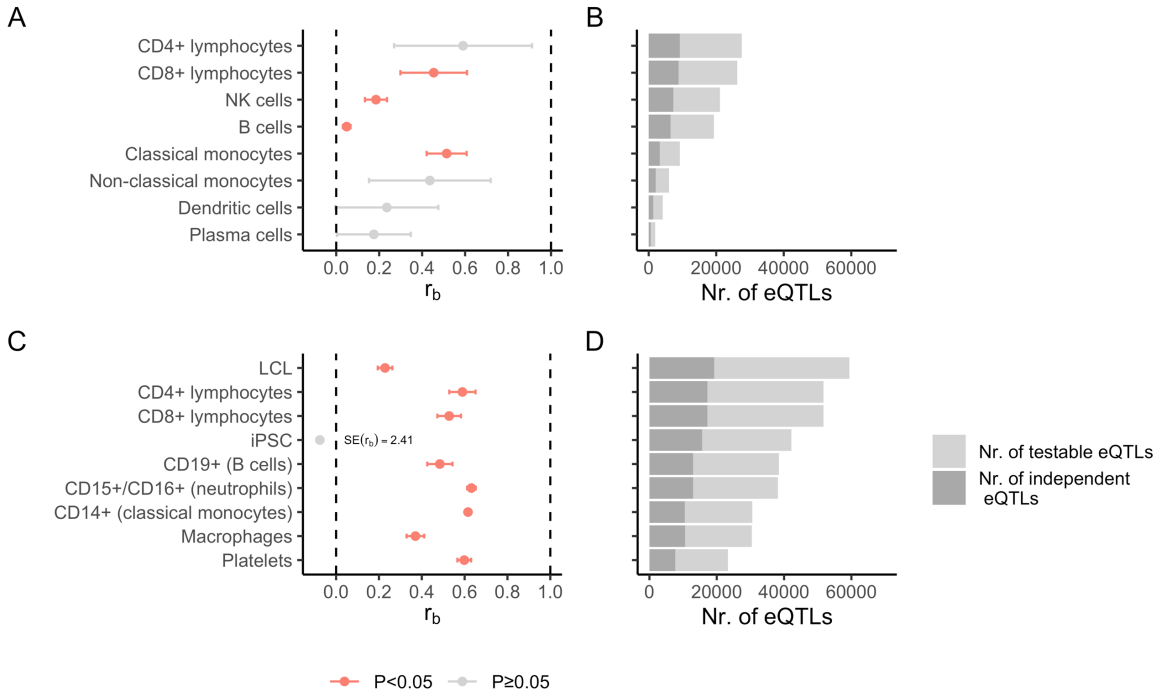

GTEx tissues:
Brain Cortex
λ = 0.992
GTEx tissues:
Brain Frontal Cortex BA9
λ = 0.964
GTEx tissues:
Brain Hippocampus
λ = 1.031
GTEx tissues:
Brain Hypothalamus
λ = 0.981
GTEx tissues:
Brain Nucleus accumbens basal ganglia
λ = 0.988
GTEx tissues:
Brain Putamen basal ganglia
λ = 0.981
GTEx tissues:
Breast Mammary Tissue
λ = 0.964
GTEx tissues:
Cells EBV-transformed lymphocytes
λ = 0.983
GTEx tissues:
Cells Transformed fibroblasts
λ = 0.987
GTEx tissues:
Colon Sigmoid
λ = 1.015
GTEx tissues:
Colon Transverse
λ = 1.033
GTEx tissues:
Esophagus Gastroesophageal Junction
λ = 1.01
− log10 (observed P -value)
− log10 (expected P -value)


GTEx tissues:
Esophagus Mucosa
λ = 0.976
GTEx tissues:
Esophagus Muscularis
λ = 1.058
GTEx tissues:
Heart Atrial Appendage
λ = 0.979
GTEx tissues:
Heart Left Ventricle
λ = 1.036
GTEx tissues:
Liver
λ = 1.037
GTEx tissues:
Lung
λ = 1.018
GTEx tissues:
Muscle Skeletal
λ = 1.084
GTEx tissues:
Nerve Tibial
λ = 1.089
GTEx tissues:
Ovary
λ = 0.981
GTEx tissues:
Pancreas
λ = 1.042
GTEx tissues:
Pituitary
λ = 1.085
GTEx tissues:
Prostate
λ = 1.003
− log10 (observed P -value)
− log10 (expected P -value)



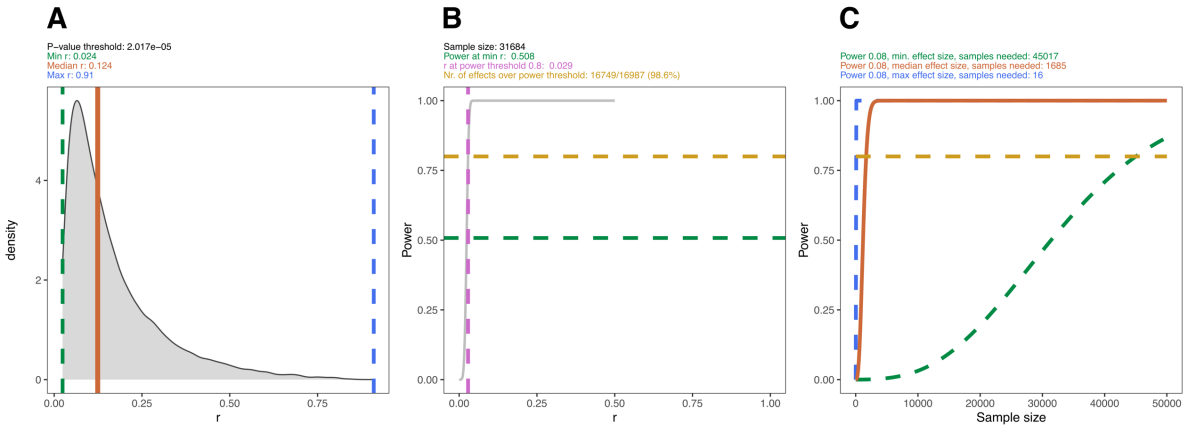
Supplementary Figure 6B. QQ-plots for investigating the inflation of eQTS signal in cell lines and bulk tissues from GTEx v7. Each plot visualises two-sided -log$_{10}$(P-values) from each replication analysis (Spearman correlation; per-dataset effects were meta-analyzed if replication included multiple datasets from the same cell- or tissue type; **Methods**, **Supplementary Methods**). Replication datasets from whole blood and from individual immune cell types are outlined with a red header. For GTEx replications, separate discovery meta-analysis was performed without GTEx, QQ-plots and lambda inflation values were calculated using eQTS effects reaching FDR<0.05 in this separate discovery analysis. Because most of the GWAS

summary statistics used in eQTS analyses originated from GWASs performed in European cohorts and discovery meta-analysis included the majority of European samples, we included only samples of European ancestry into GTEx replication analyses shown here.

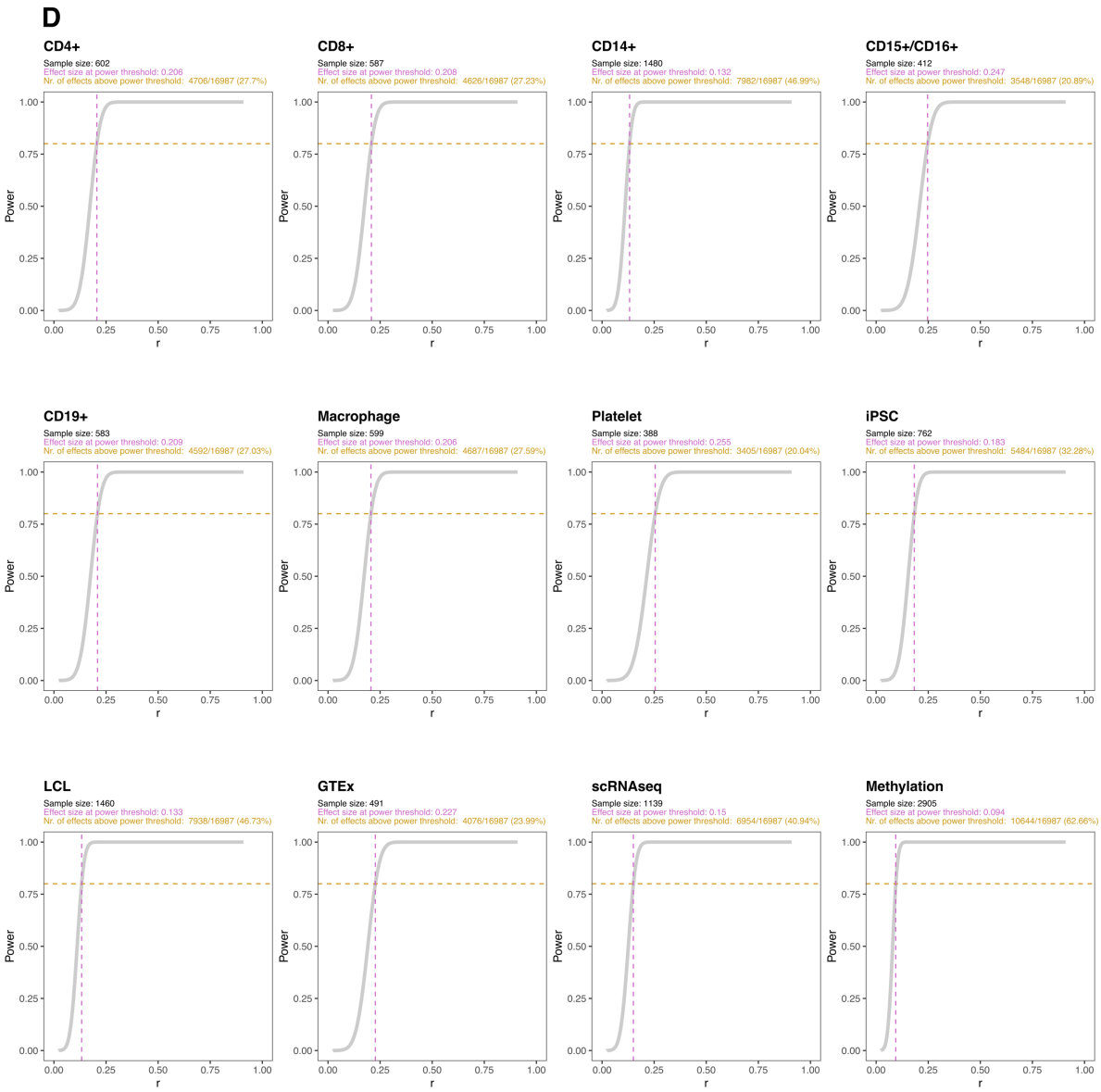


Supplementary Figure 7. Correlation analysis (r_b) for *trans*-eQTL replication in scRNA-seq data, cell lines and purified cell types. The r_b statistic is a measure for correlation of effect sizes; higher r_b values indicate stronger sharing of *trans*-eQTL signal between studies. A. r_b analysis in scRNA-seq data. Dot indicates the r_b, error bar indicates the standard error (SE) of r_b. In order to calculate P-value, r_b and SE(r_b) were converted to Z-score and P-value was derived from chi-squared distribution with one degree of freedom. B. Number of testable *trans*-eQTLs in each scRNA-seq cell type. The length of the bars indicates the number of all testable *trans*-eQTLs and dark grey bars indicate the subset of independent effects (SNPs ±1 Mb from lead discovery SNP removed) which were used for r_b calculation. C. r_b analysis in purified cell types and cell lines. D. Number of testable *trans*-eQTLs in each purified cell type and cell line.

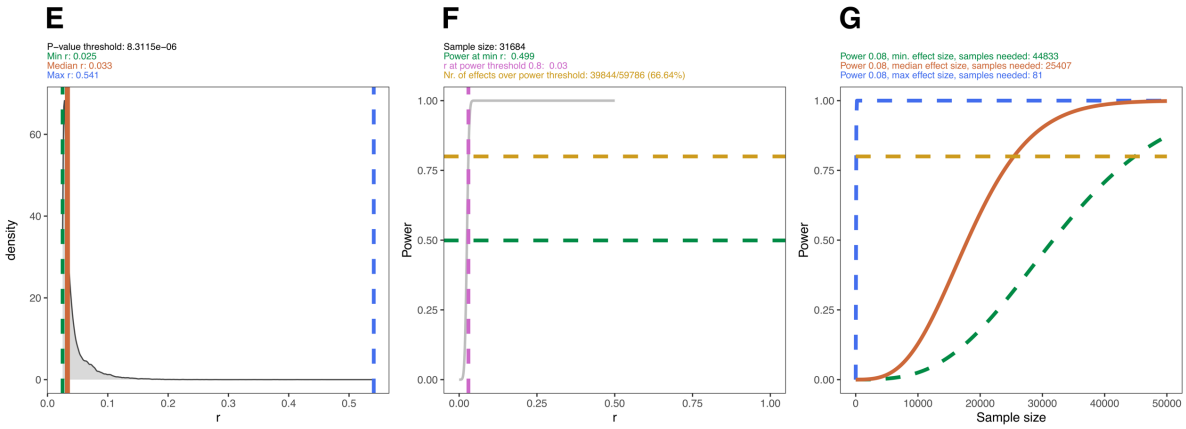
cis–eQTLs: discovery meta-analysis



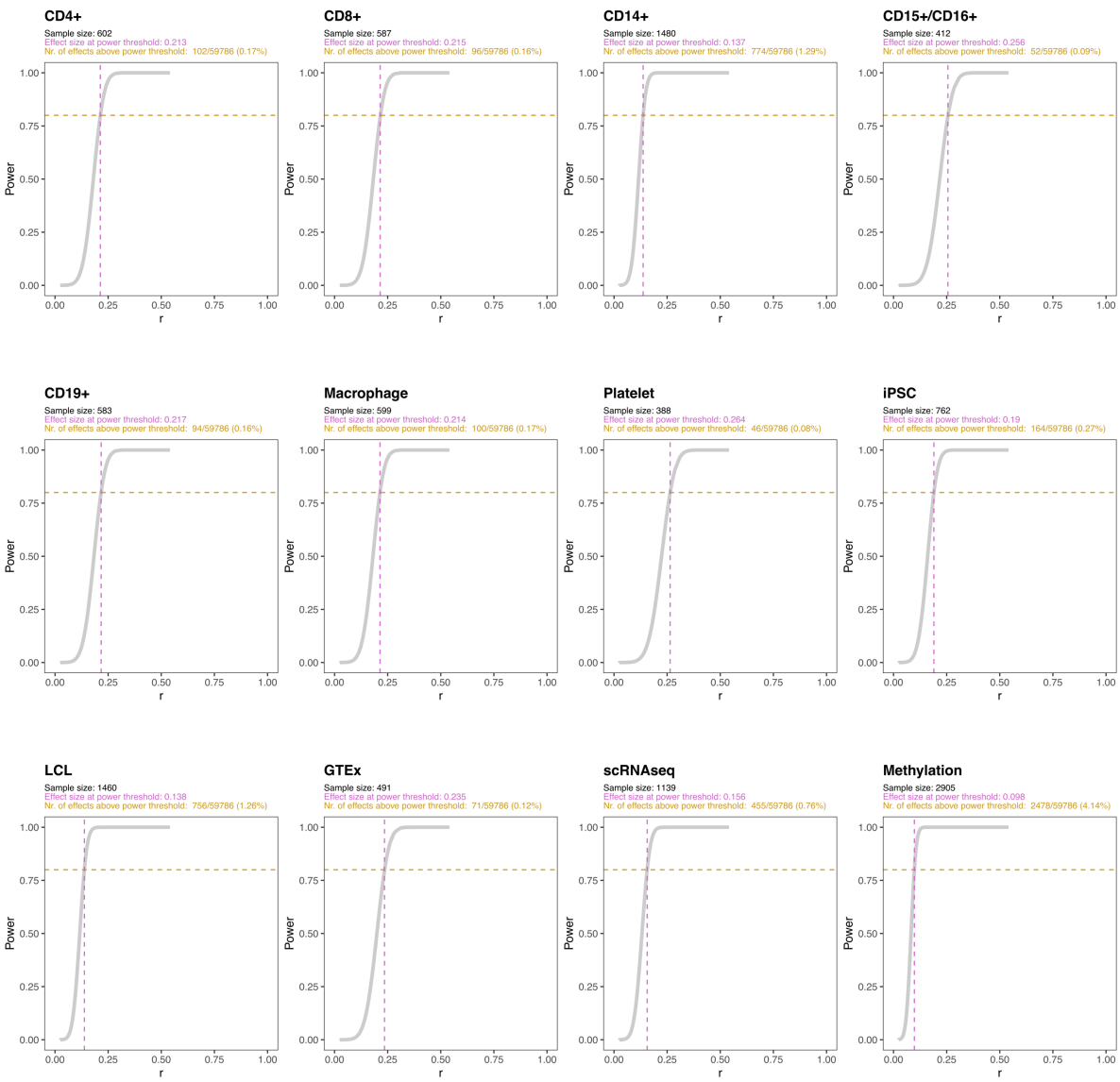
cis–eQTLs: replication datasets



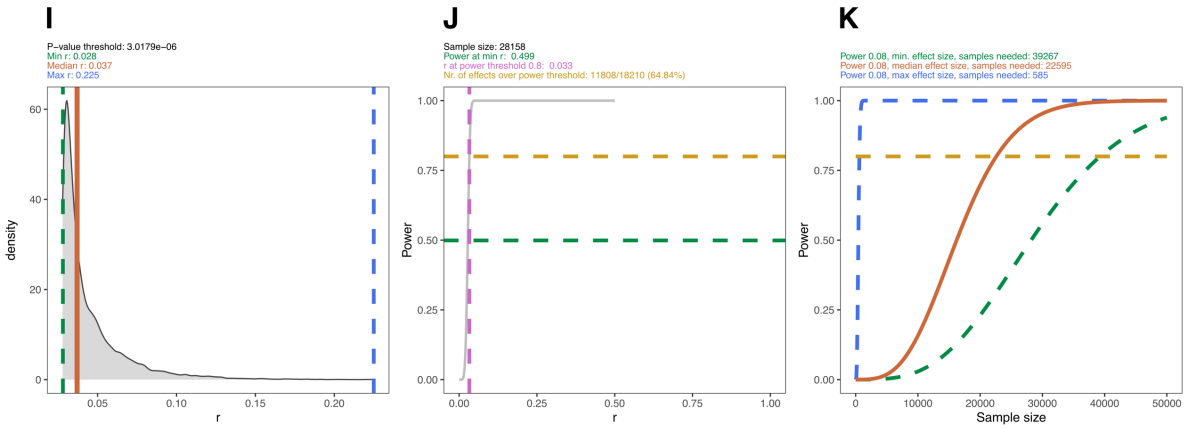
trans–eQTLs: discovery meta-analysis



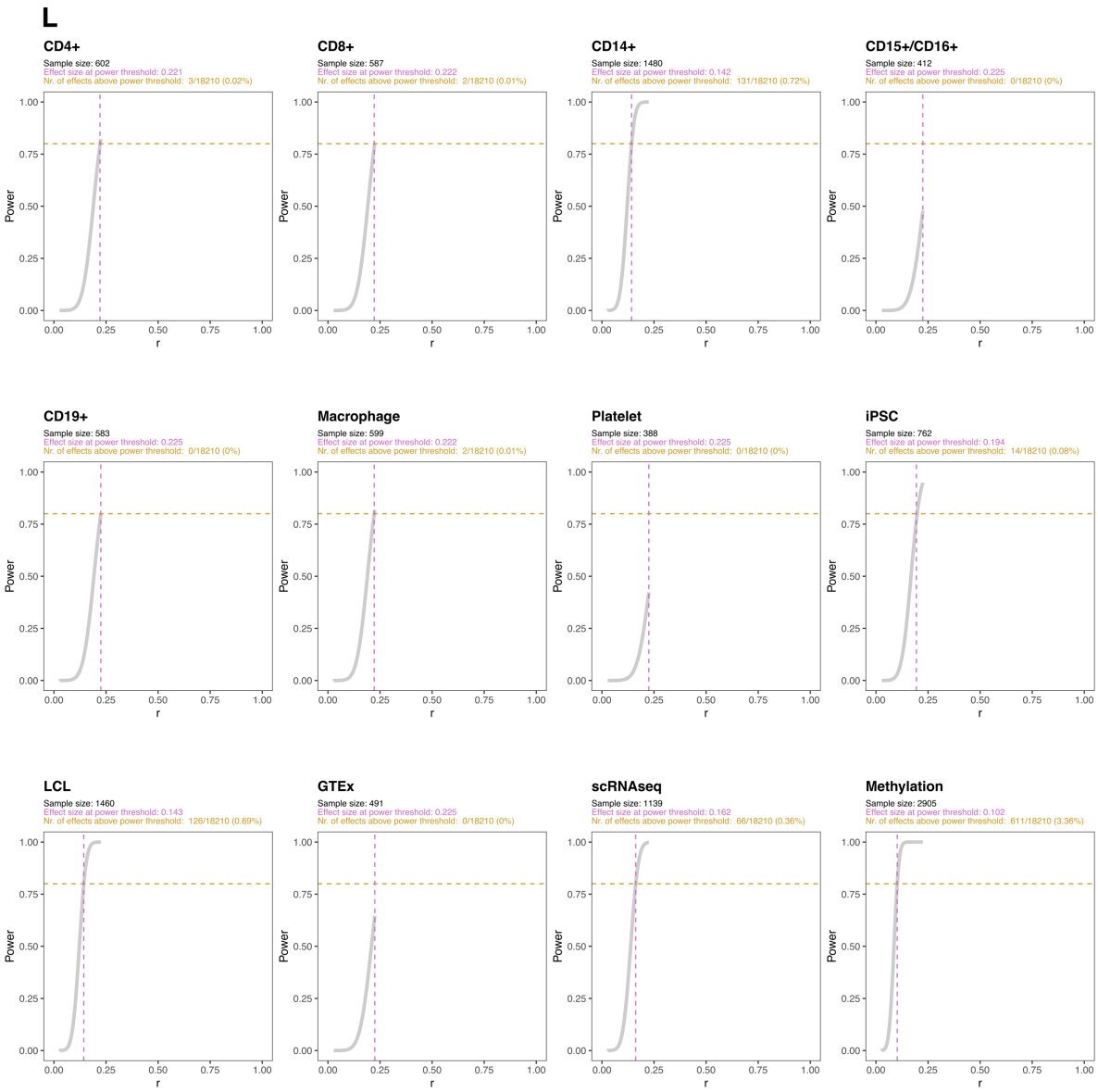
trans–eQTLs: replication datasets



eQTSs: discovery meta-analysis

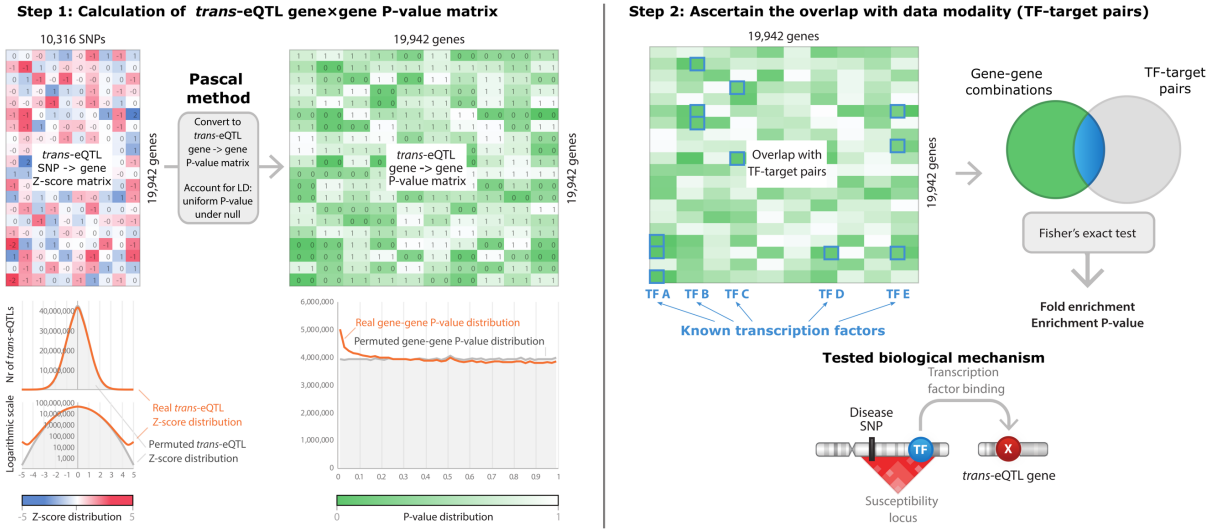


eQTSs: replication datasets

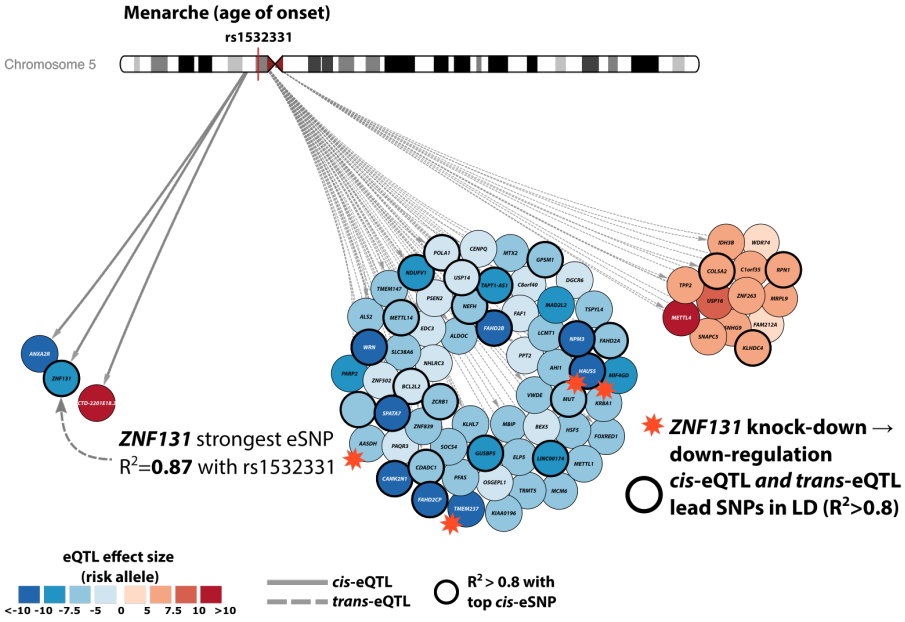


Supplementary Figure 8. Distribution of significant (FDR<0.05) *cis*-eQTL, *trans*-eQTL and eQTS effect sizes (absolute correlation coefficients r, calculated from meta-analysis Z-scores) and the replication power of replication datasets. A, E, I: Effect size distributions and corresponding minimal, median and maximal effect sizes. B, F, J: Relationship between discovery analyses effect sizes and analysis power. Outlined are power to detect minimal discovery effect size (green), effect size corresponding to power level of 0.8 (purple), and line signifying the power threshold of 0.8 (yellow). C, G, K: Relationship between sample size and power in the discovery analysis. Different lines indicate how many samples would be needed to detect effects with minimal effect size (green), median effect size (orange) and maximal effect size (yellow). D, H, L: Relationship between replication dataset effect size and replication power. Lines indicate the power threshold of 0.8 and corresponding replication effect threshold (purple) for given dataset. For GTEx v7, largest sample size over all the tissues was used for those power estimations, real replication power in remaining tissues is smaller. BIOS methylation dataset contains samples which were part of the discovery meta-analysis, therefore this should not be considered as independent replication.

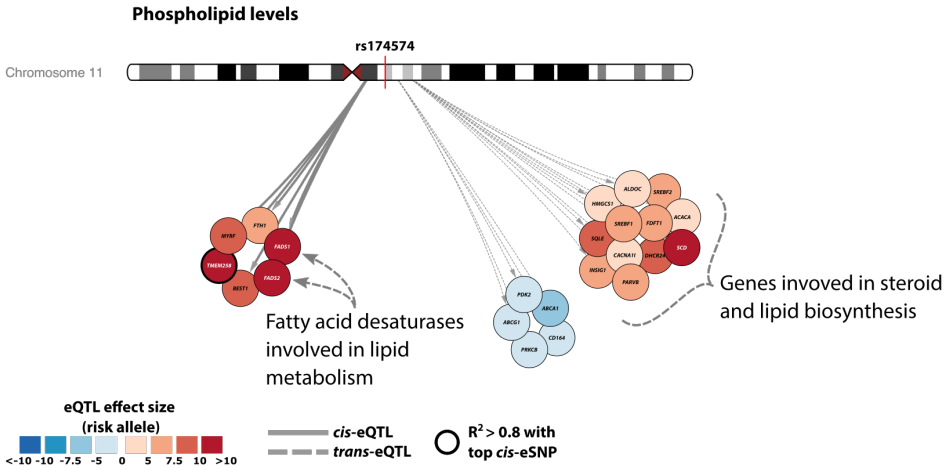
Supplementary Figure 9. Putative mechanisms leading to *trans*-eQTL effects. We explored putative mechanisms that might explain the *trans*-eQTLs observed in discovery meta-analysis by series of enrichment analyses (two-sided Fisher's exact test). To perform these enrichment analyses, we first converted *trans*-eQTL results to a *cis-trans* gene-gene matrix using the Pascal method.



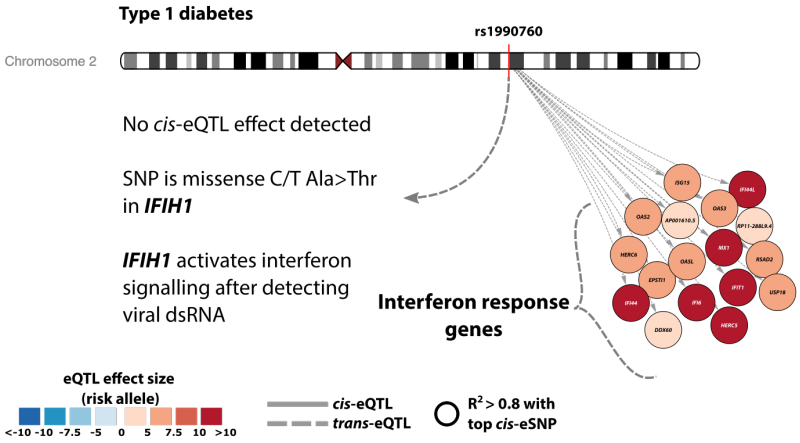
Supplementary Figure 10. Strategy of converting *trans*-eQTL data to *cis-trans* gene-gene matrix by Pascal method. Using the permuted Z-score summary statistics of our *trans*-eQTL analysis, we corrected the *trans*-eQTL Z-score statistics for LD between variants and created gene-gene P-value matrices (left) for the unpermuted and permuted data.



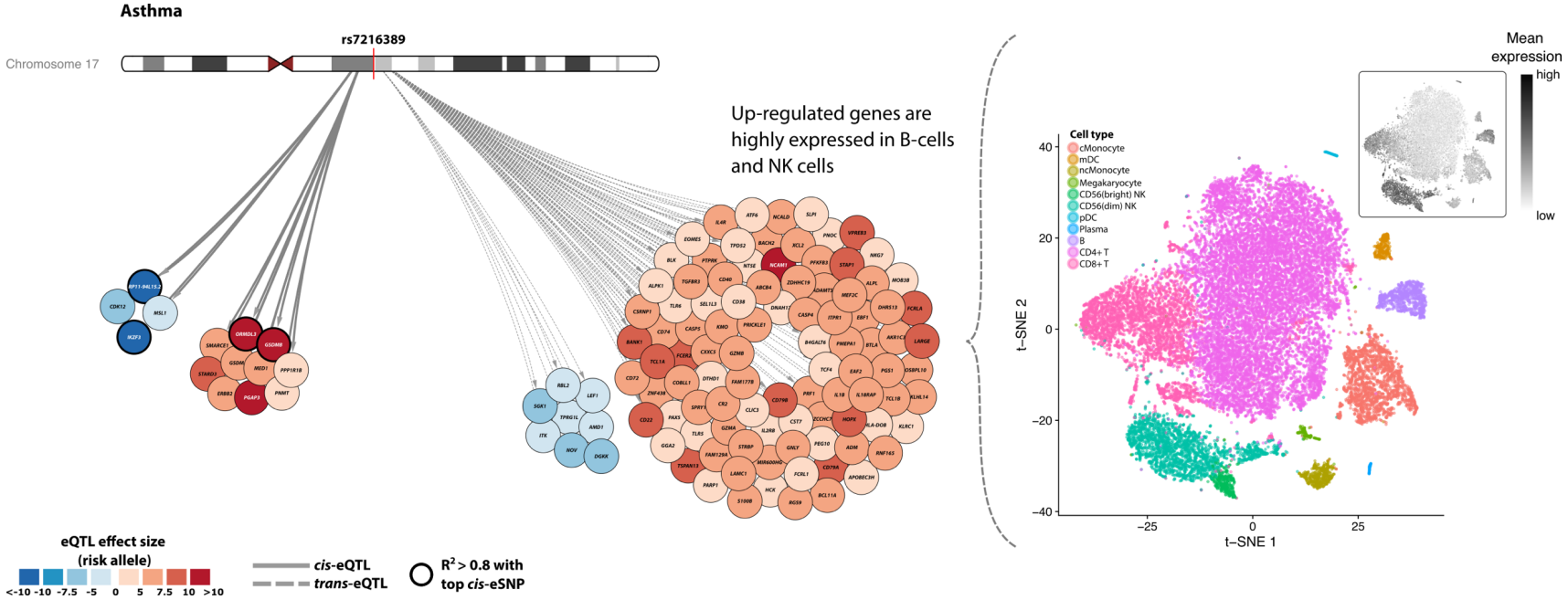
Supplementary Figure 11A. Age-of-menarche-associated SNP rs1532331 (5p12) is in high LD (R^2>0.8, 1kG p1v3 EUR) with the lead *cis*-eQTL effect for *ZNF131* which encodes transcription factor. In total, we identified 3 *cis*-eQTL genes and 75 *trans*-eQTL genes associated with this variant. In a recent short hairpin RNA knockdown experiment of *ZNF131*, three separate cell isolates showed downregulation of four genes that we identified as *trans*-eQTL genes: *HAUS5*, *TMEM237*, *MIF4GD* and *AASDH* (indicated by red stars). The product of *ZNF131* has been hypothesized to inhibit estrogen signaling, which may explain how the SNP in this locus contributes to altering the age of menarche. For *trans*-eQTLs, the outline indicates that lead *cis*-eQTL from the full discovery meta-analysis and lead *trans*-eQTL SNPs (from locus-wide *trans*-eQTL analysis in the subset of 4,339 samples) are in high LD (R^2>0.8, 1kG p1v3 EUR), suggesting co-localization.



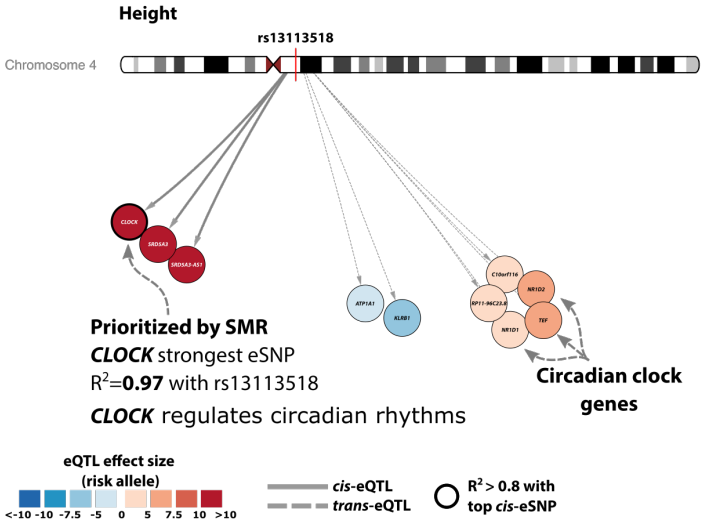
Supplementary Figure 11B. *Trans*-eQTLs extend insight for loci with multiple *cis*-eQTL effects. In the *FADS1*/*FADS2* locus, rs174574 (11q12.2) is associated with lipid levels and affects 6 genes in *cis* and 17 genes in *trans*. The strongest *cis*-eQTLs in this locus modulate the expression of *FADS1*, *FADS2* and *TMEM258*, with the latter being in high LD with the GWAS SNP (R^2>0.8, 1kG p1v3 EUR). From those genes, *FADS1* and *FADS2* have been implicated to affect lipid levels since these encode fatty acid desaturases. Consistent with their function, *trans*-eQTL genes from this locus are highly enriched for triglyceride metabolism in REACTOME ($P<4.1\times10^{-9}$, GeneNetwork pathway enrichment method which leverages large-scale gene expression data; www.genenetwork.nl).



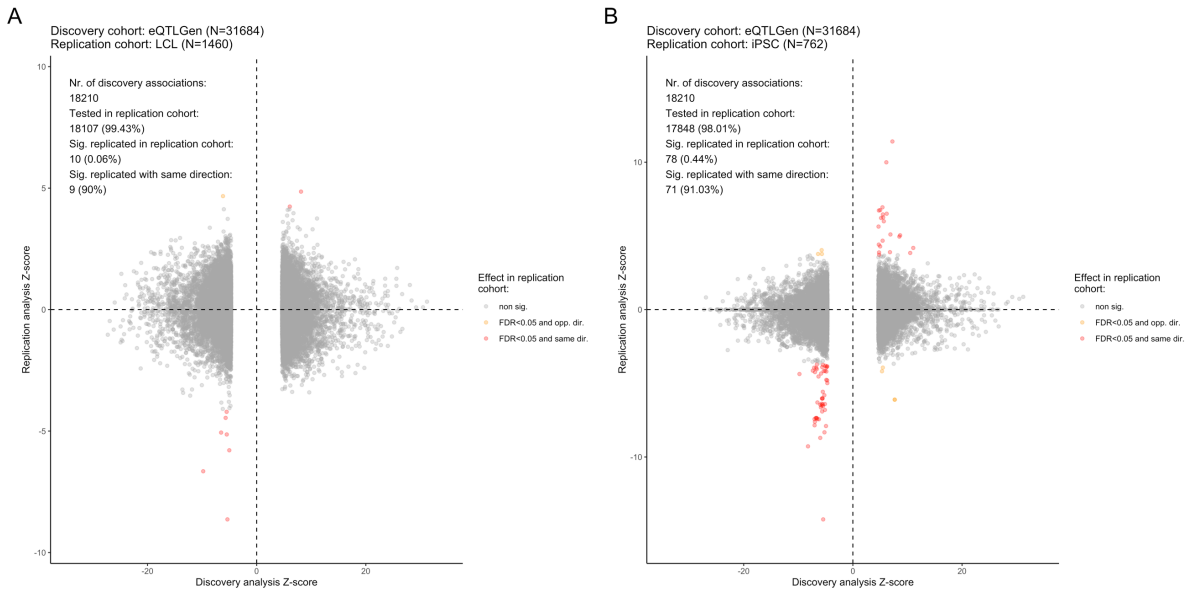
Supplementary Figure 11C. *Trans*-eQTLs can shed light on loci with no detectable *cis*-eQTLs. rs1990760 (2q24.2) is associated with multiple immune-related traits (type 1 diabetes (T1D), inflammatory bowel disease (IBD), systemic lupus erythematosis (SLE) and psoriasis). For this SNP there are 17 significant (FDR<0.05) *trans*-eQTL effects, but no detectable gene-level *cis*-eQTL. The risk allele for this SNP causes an Ala946Thr amino acid change in the RIG-1 regulatory domain of MDA5 (encoded by gene *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 signaling 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.



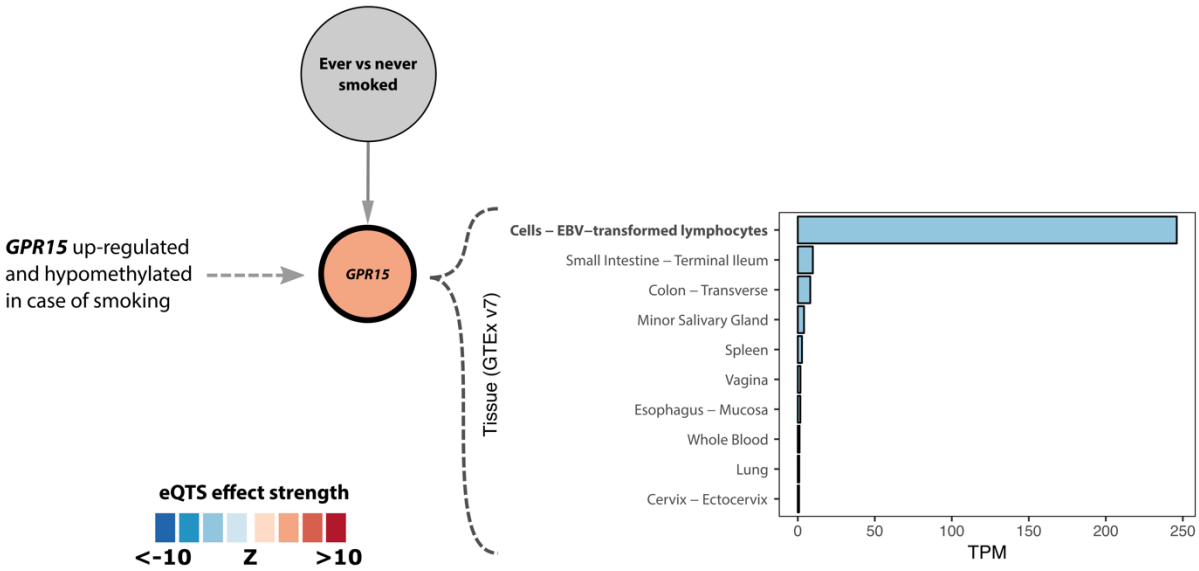
Supplementary Figure 11D. *Trans*-eQTLs can reveal cell type composition effects of the trait-associated SNP. Asthma-associated SNP rs7216389 (17q12-q21.1) has 14 *cis*-eQTL effects (FDR<0.05), most notably on *IKZF3*, *GSDMB*, and *ORMDL3* (left). Ninety-four out of the 104 *trans*-eQTL genes were up-regulated by the risk allele for asthma and were mostly expressed in B cells and natural killer cells (right). *IKZF3* encodes the protein which is part of the Ikaros transcription factor family that regulates B-cell proliferation, suggesting that a decrease of the product of *IKZF3* leads to an increased number of B cells and resulting *trans*-eQTL effects caused by cell-type composition differences.



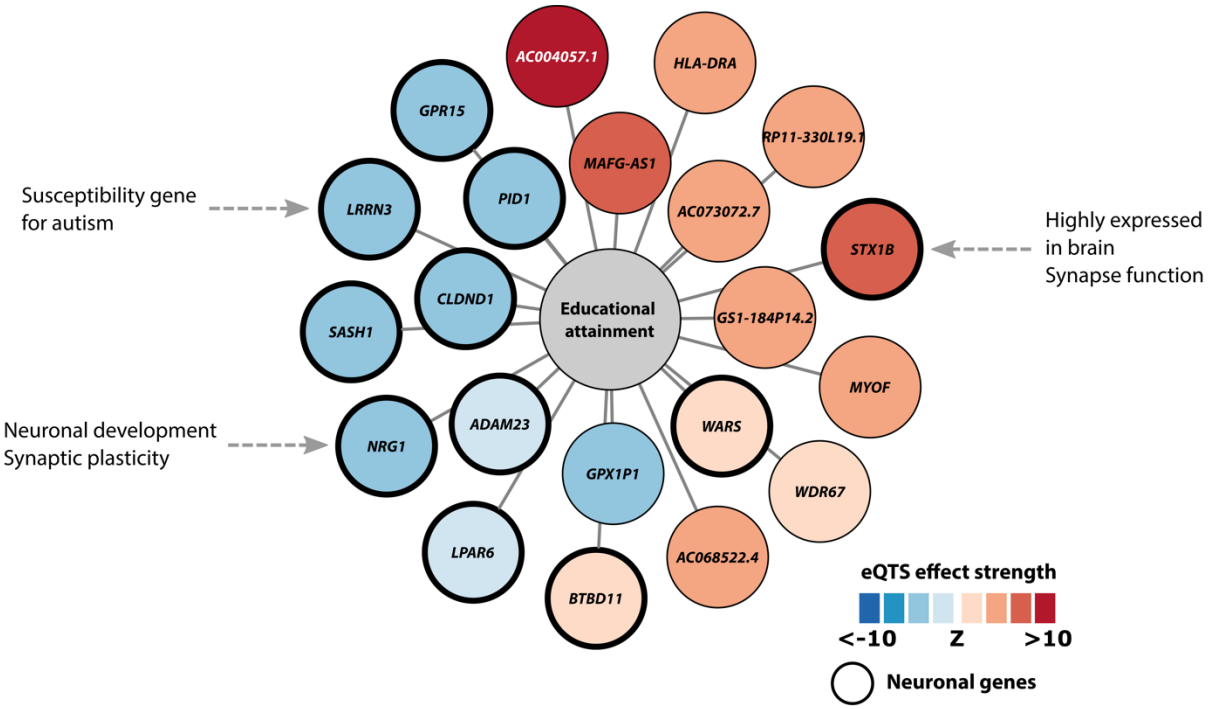
Supplementary Figure 11E. *Trans*-eQTLs can suggest the involvement of pathways that are not previously associated with certain complex traits. Height-associated SNP rs1311351834 (4q12) is in high LD (R^2>0.8, 1kG p1v3 EUR) with the lead *cis*-eQTL SNP for the *CLOCK* gene. The upregulated TF CLOCK forms a heterodimer with TF BMAL1, and the resulting protein complex regulates circadian rhythm. 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.



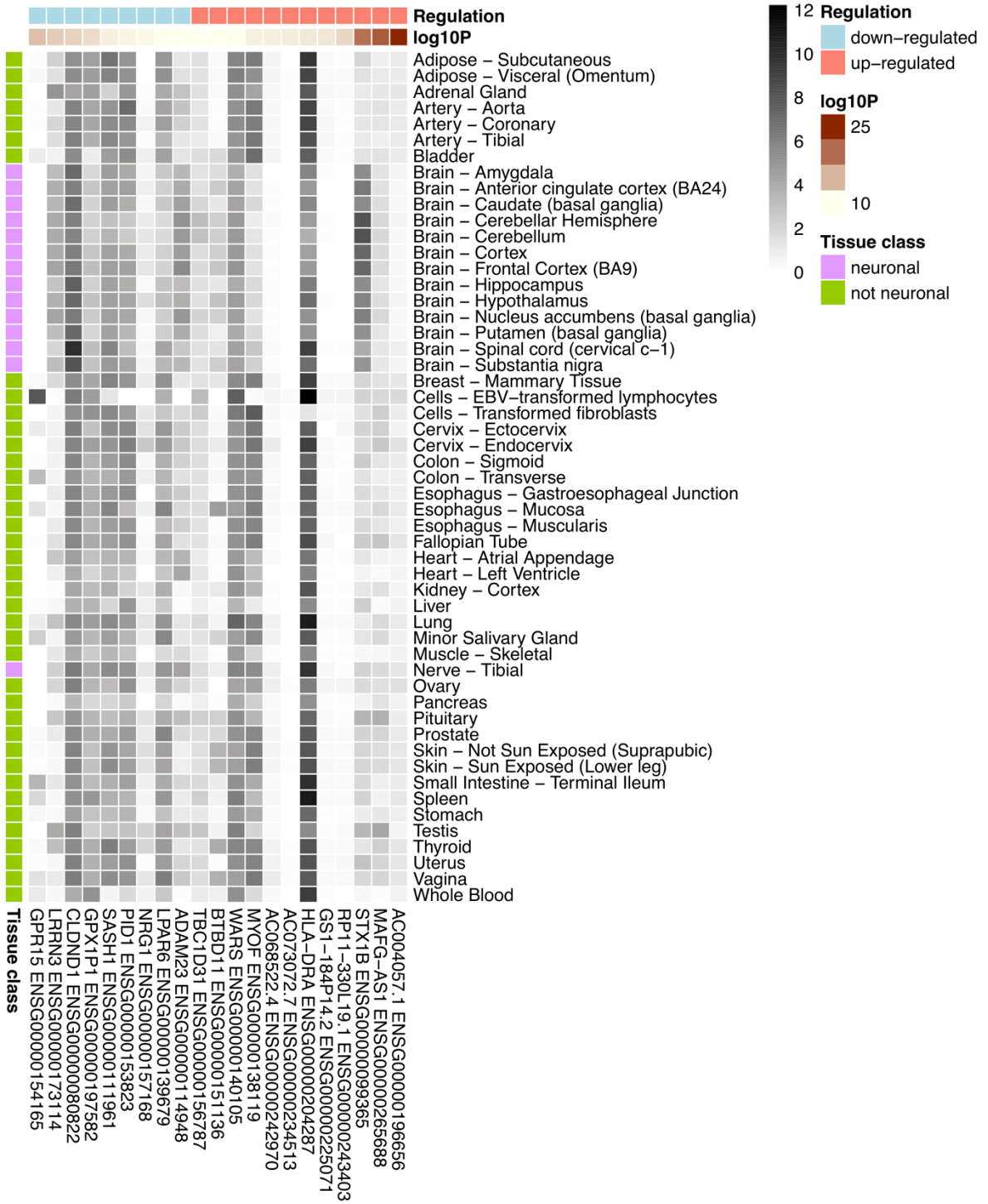
Supplementary Figure 12. Replication analyses for eQTSs. Comparison of eQTS discovery meta-analysis Z-scores (x-axis) and replication analysis Z-scores (y-axis) in LCL (A) and iPSC (B) cell lines. Note, that for better visualization, the scale of the y-axis varies on each plot. Grey dots indicate eQTS effects that were not significant in the replication study, red dots indicate significant (Benjamini-Hochberg FDR<0.05) eQTS effects with identical direction with the discovery, and orange dots indicate significant effects with opposite direction with the discovery.



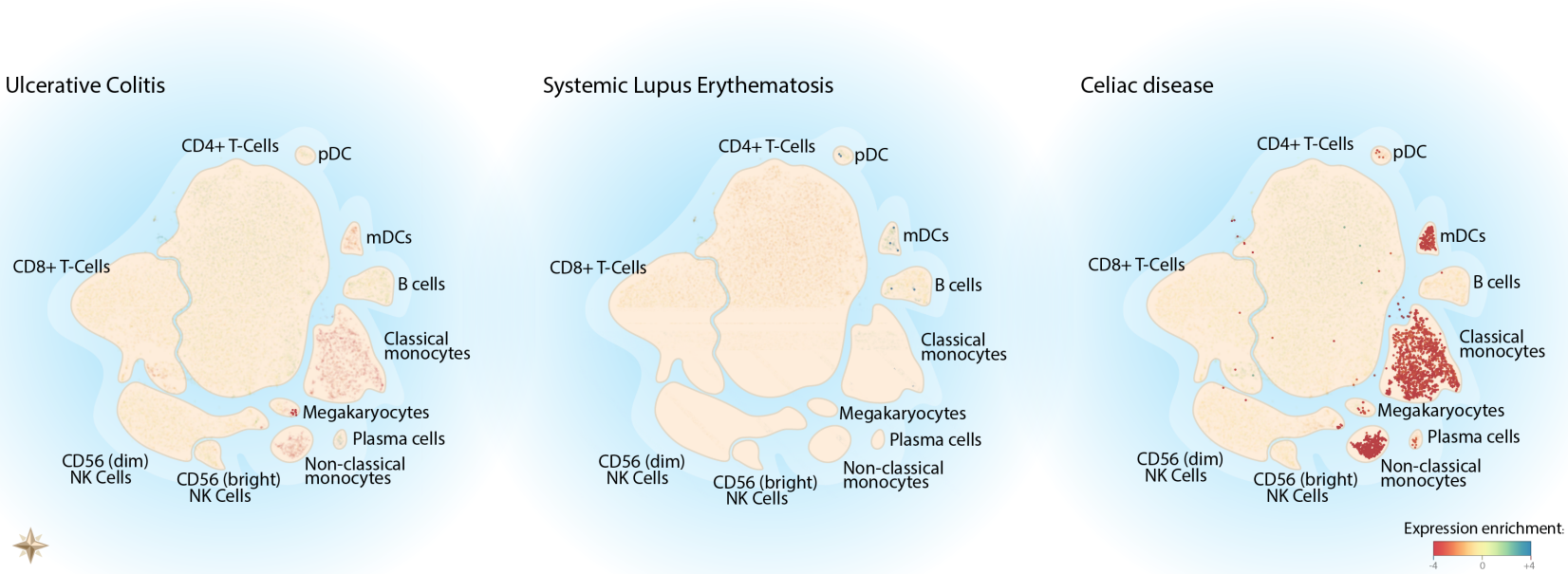
Supplementary Figure 13A. eQTS analysis can identify genes relevant for non-blood traits. For example, the expression of *GPR15* (discovery eQTS meta-analysis $P=3.7\times10^{-8}$, FDR=0.00137) is associated with the trait 'ever versus never smoking'. *GPR15* is a biomarker for smoking that is overexpressed and hypomethylated in smokers. We observe strong *GPR15* expression in lymphocytes, suggesting that the association with smoking could originate from a change in the proportion of T cells in blood. As *GPR15* is involved in T cell homing and has been linked to colitis and inflammatory phenotypes, it is hypothesized to be involved in the systematic inflammation induced by tobacco smoking. The expression of *GPR15* in GTEx v7 tissues is visualized in the right pane (bars indicate transcripts per million (TPM), 10 tissues with the highest *GPR15* expression are visualized).



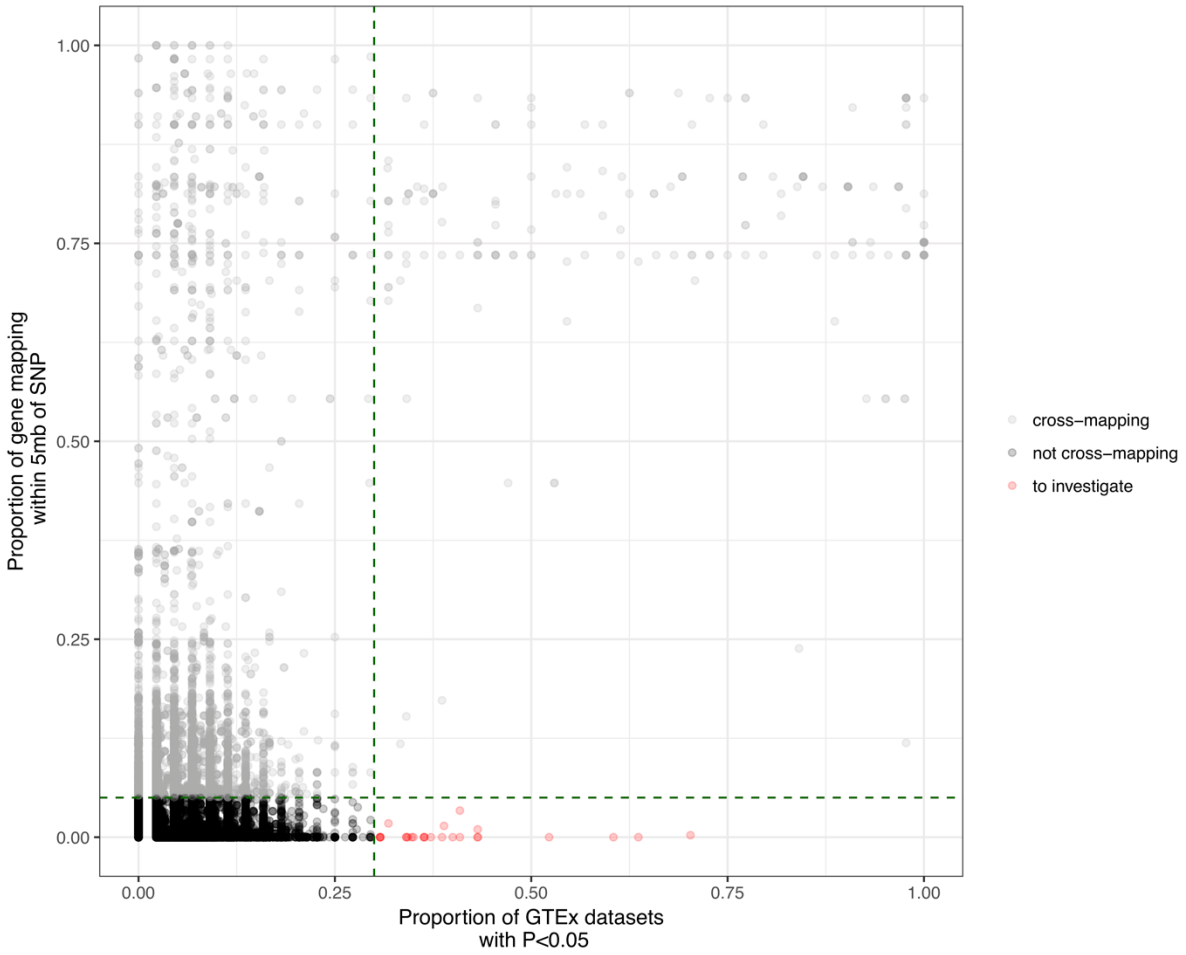
Supplementary Figure 13B. The PGS for educational attainment correlated significantly with the expression of 21 genes (FDR<0.05). Several of the strongly associated genes are known to be involved in neuronal processes (thick border).



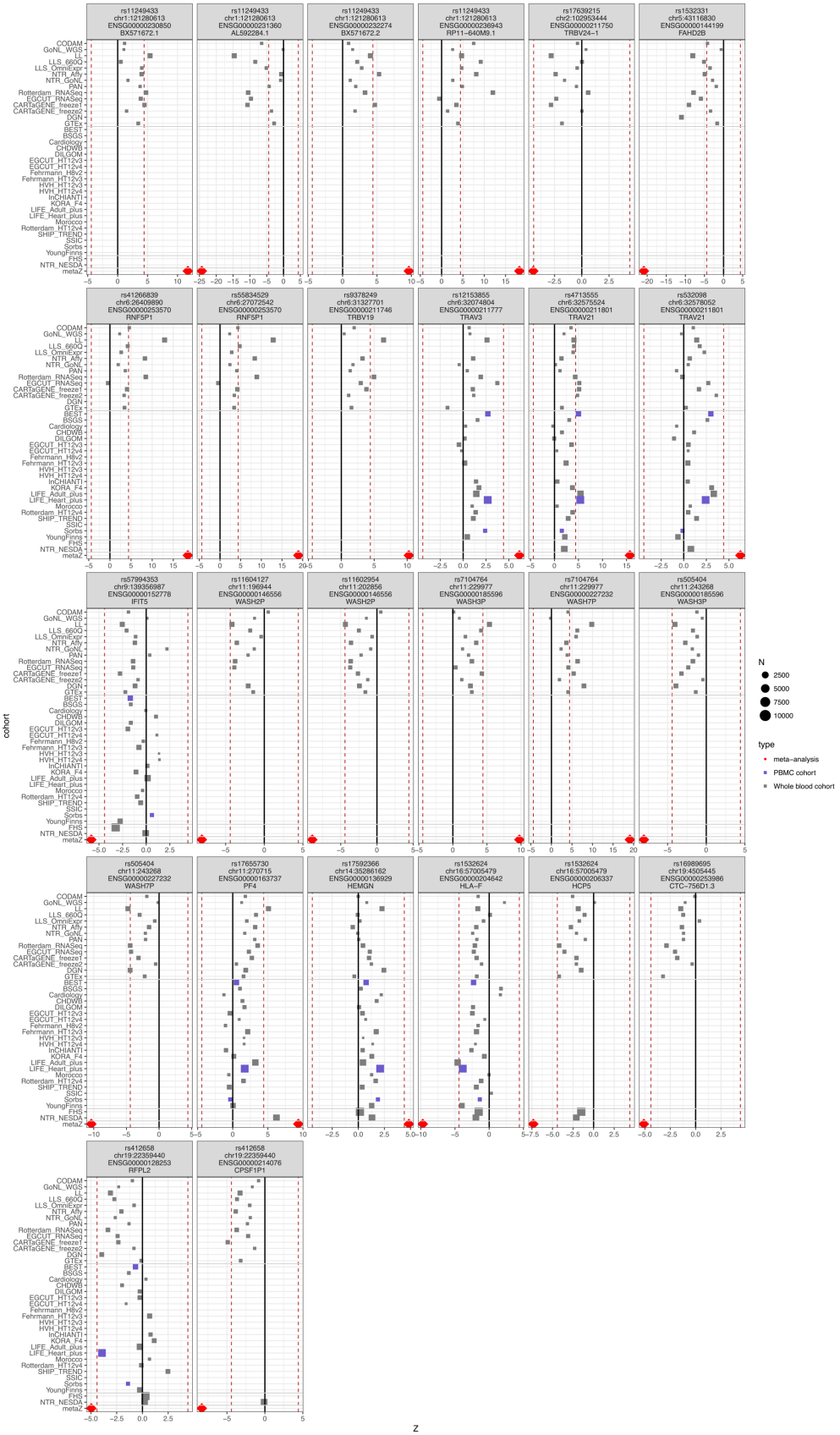
Supplementary Figure 13C. Expression levels of significant eQTS genes for educational attainment in GTEx v7 tissues. The log_2-scaled expression values in transcripts per million (log_2(TMP+1)) are visualised on the heatmap. The regulation bar indicates if the gene is up- (red) or down-regulated (blue) relative to higher PGS for educational attainment; the log10P bar shows the discovery eQTS meta-analysis association significance for a given gene, represented as -log_{10}(P); the tissue class column indicates whether the GTEx tissue is neuronal. Several of the eQTS genes for educational attainment show expression in brain and neuronal tissues



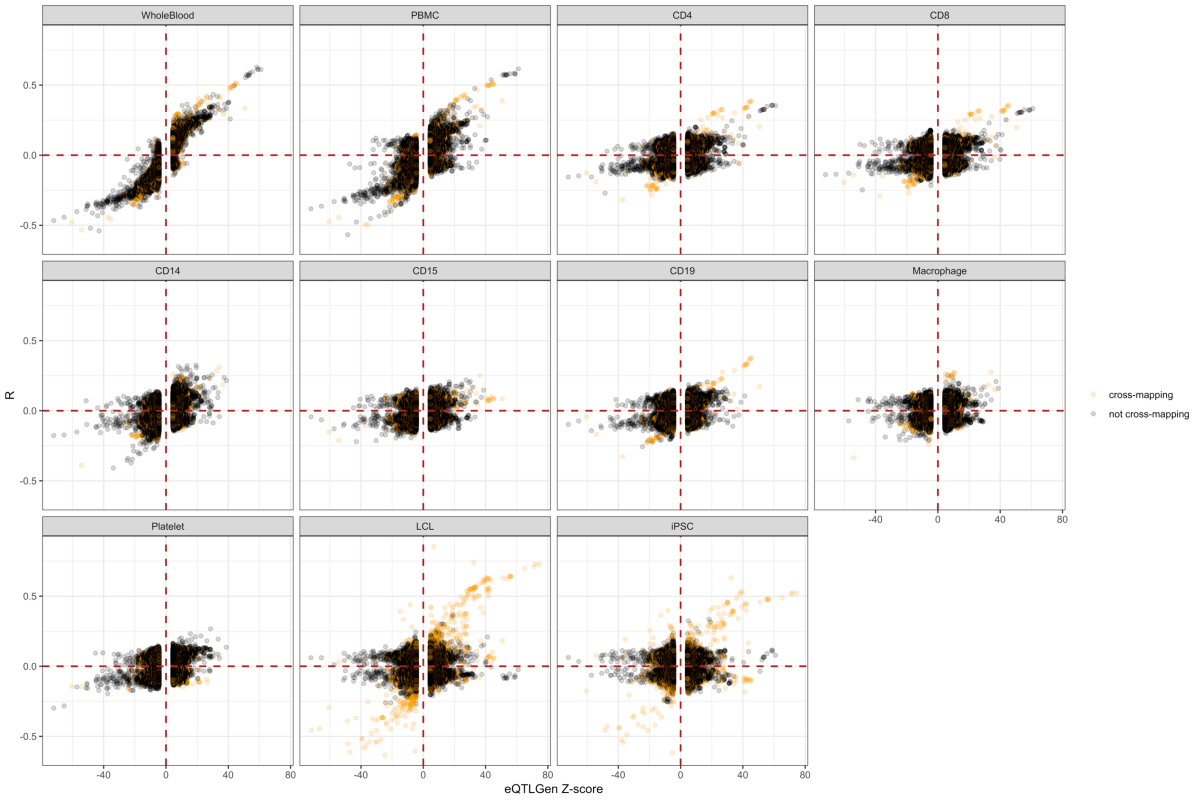
Supplementary Figure 13D. Cell types associated with autoimmune disorders. eQTS gene signatures in scRNA-seq data of ~25,000 peripheral blood mononuclear cells isolated from 45 individuals (Van der Wijst et al, Nat. Genet. 2018) shows the cell types in which the genes associated with the PGS for ulcerative colitis, systemic lupus erythematosus and celiac disease are most likely expressed.



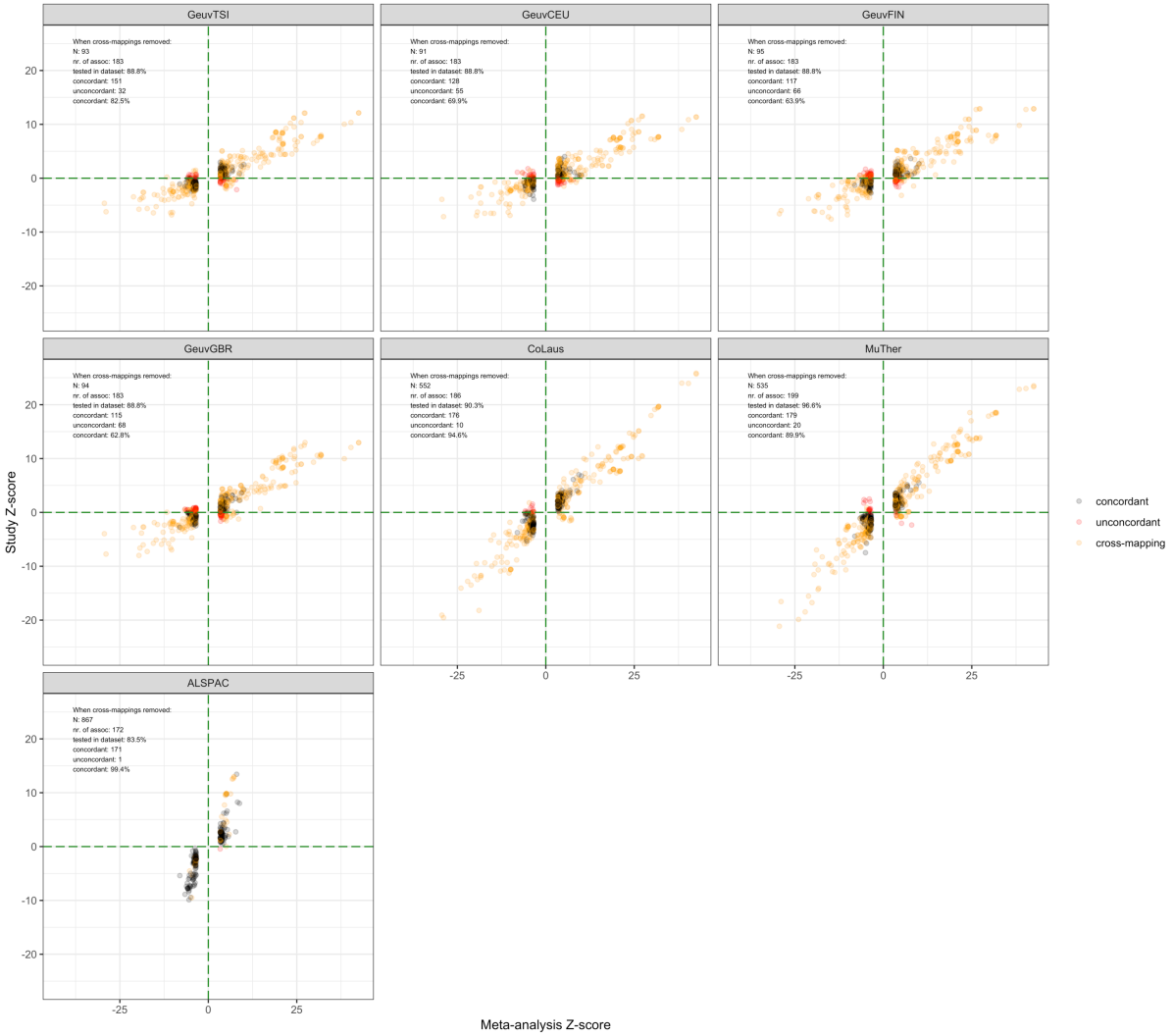
Supplementary Figure 14A. Cleaning of cross-mapping *trans*-eQTLs, based on nominal replication in GTEx v6p tissues. For that, each *trans*-eQTL gene was divided to synthetic reads and mapped against each 10 Mb region, centered around each *trans*-eQTL SNP. *Trans*-eQTLs for which >5% of all reads mapped near the *trans*-eQTL SNP were considered as potential artifacts. On the y-axis is shown the 5% cross-mapping threshold we used to filter out *trans*-eQTLs which might be caused by a read cross-mapping within the *cis* region. Red dots outline 26 *trans*-eQTLs which showed high nominal replication rates in GTEx tissues (uncorrected two-sided P<0.05; Spearman correlation) but low cross-mapping rate, and which were subsequently selected for further investigation.



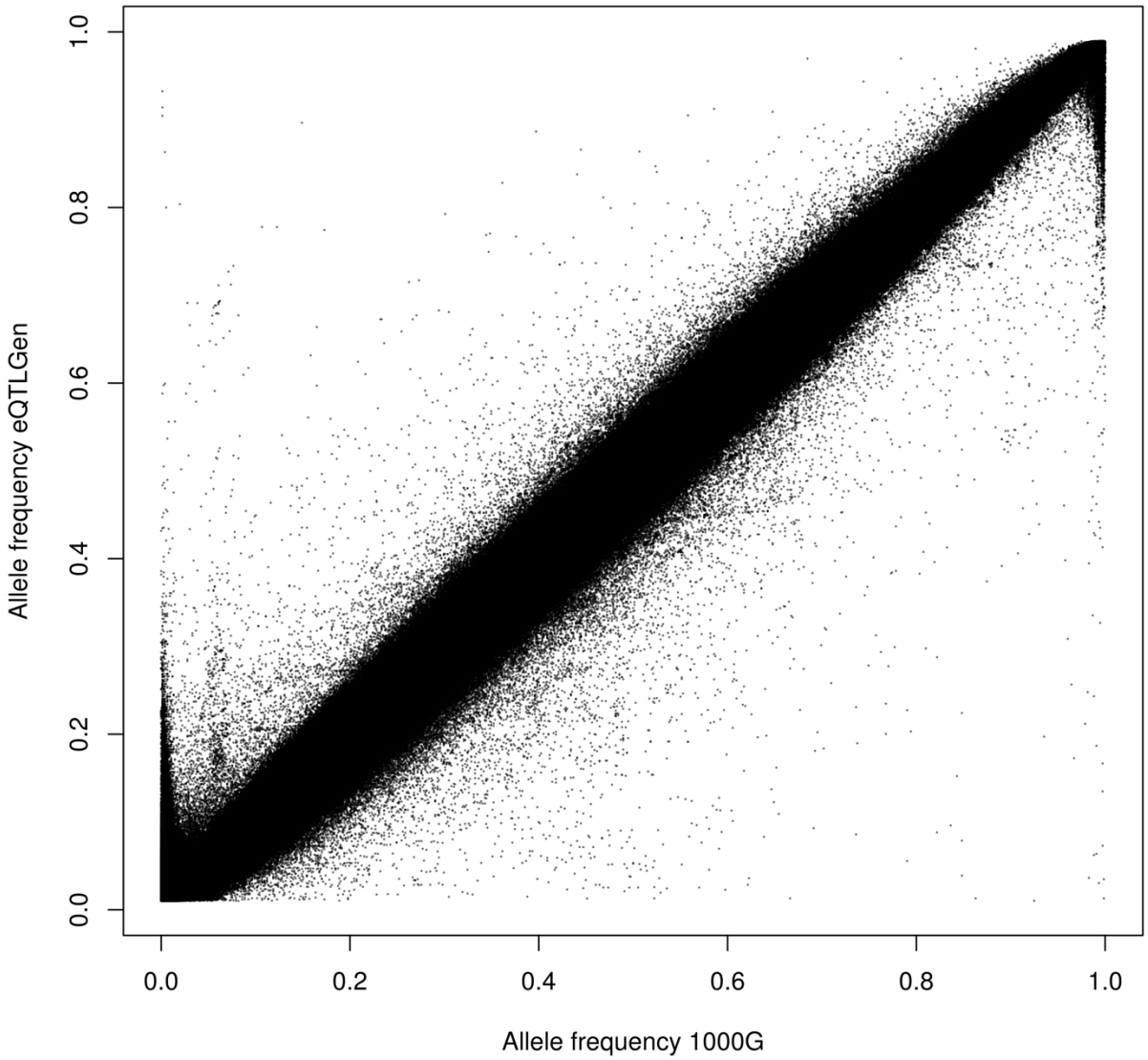
Supplementary Figure 14B. Twenty-six *trans*-eQTLs that show high replication rates in GTEx v6p tissues, yet low cross-mapping with cis region surrounding the *trans*-eQTL SNP. Forest plots show the effect directions (Z-scores) in all 37 datasets.



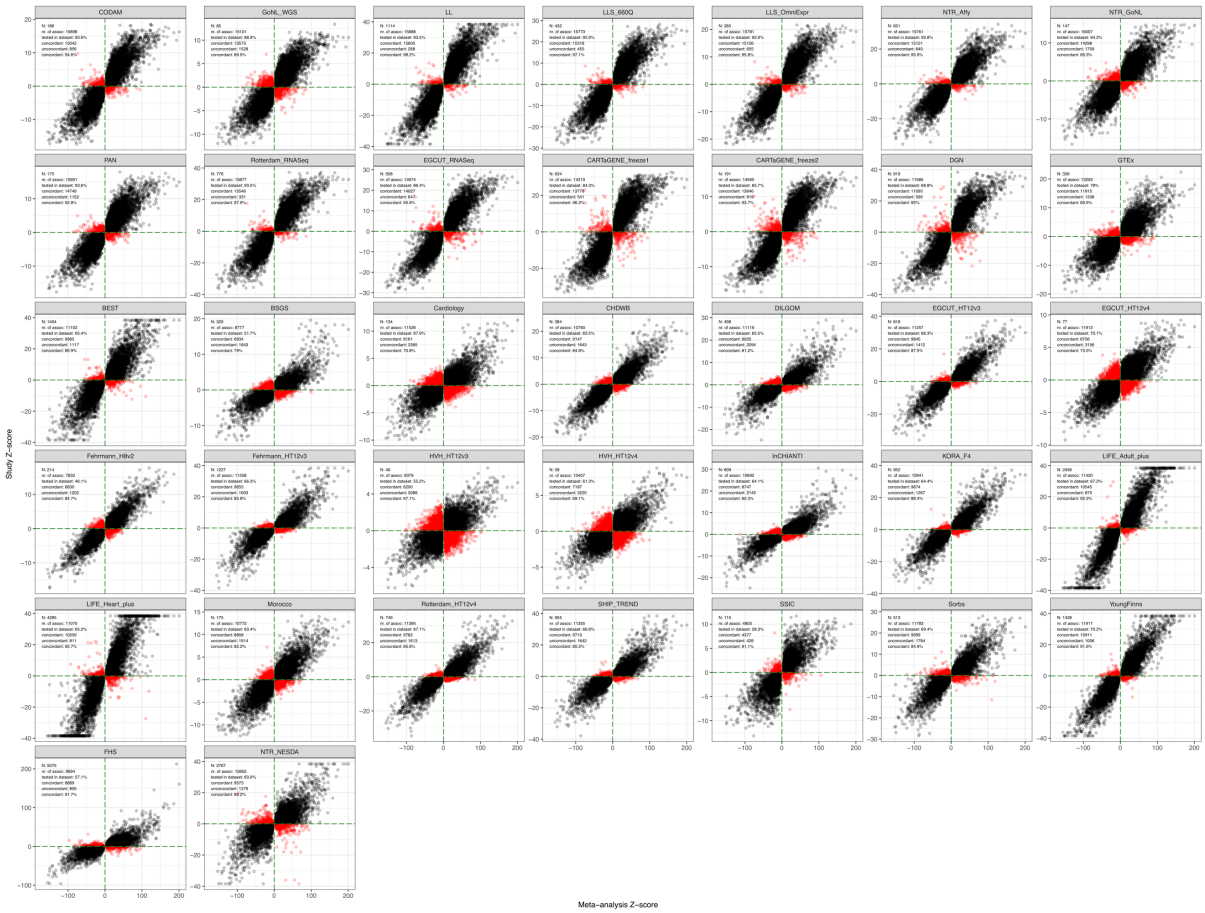
Supplementary Figure 14C. *Trans*-eQTL replication plots with cross-mapping *trans*-eQTLs outlined (yellow dots), comparing Z-scores between the discovery meta-analysis on the x-axis, and the replication datasets on the y-axis. We observed generally higher effect sizes for cross-mapping *trans*-eQTLs in several replication cell types. WholeBlood and PBMC indicate results from size-matched whole blood and PBMC subsets of the discovery analysis. Note that for better visualization, scales of y-axis differ from scales of x-axis.



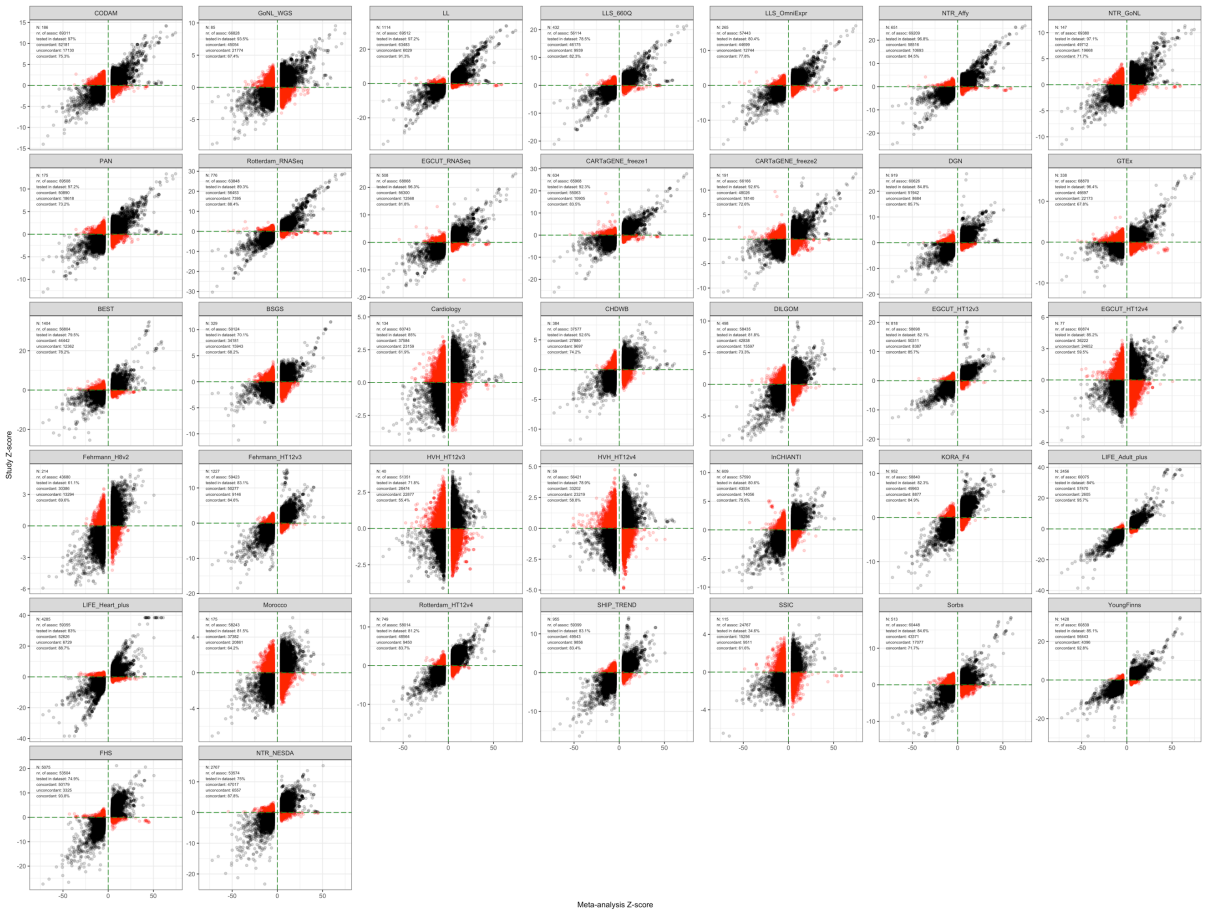
Supplementary Figure 14D. *Trans*-eQTL replication analysis in LCLs, with cross-mapping effects outlined. On the x-axis is Z-score from LCL replication meta-analysis, on the y-axis are Z-scores from individual LCL cohorts. Cross-mapping trans-eQTLs are outlined (yellow dots). Note that for better visualization, scales of y-axis differ from scales of x-axis.



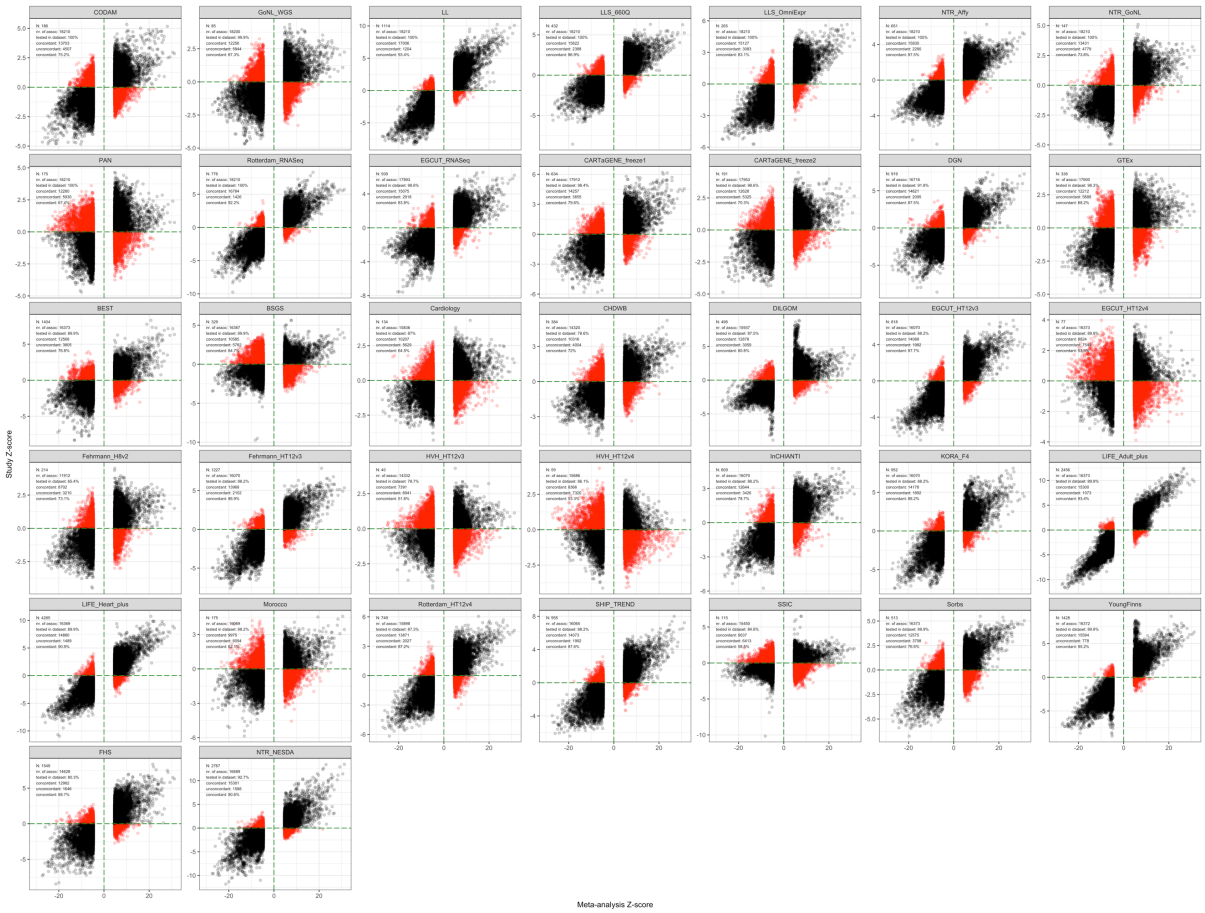
Supplementary Figure 15. SNP allele frequency (AF) comparison between eQTLGen and 1000G p1v3 EUR reference panel. eQTLGen AF calculations did not include Framingham Heart Study (N=5,075). Allele frequencies in eQTLGen are highly consistent with the 1000G p1v3 EUR reference panel.



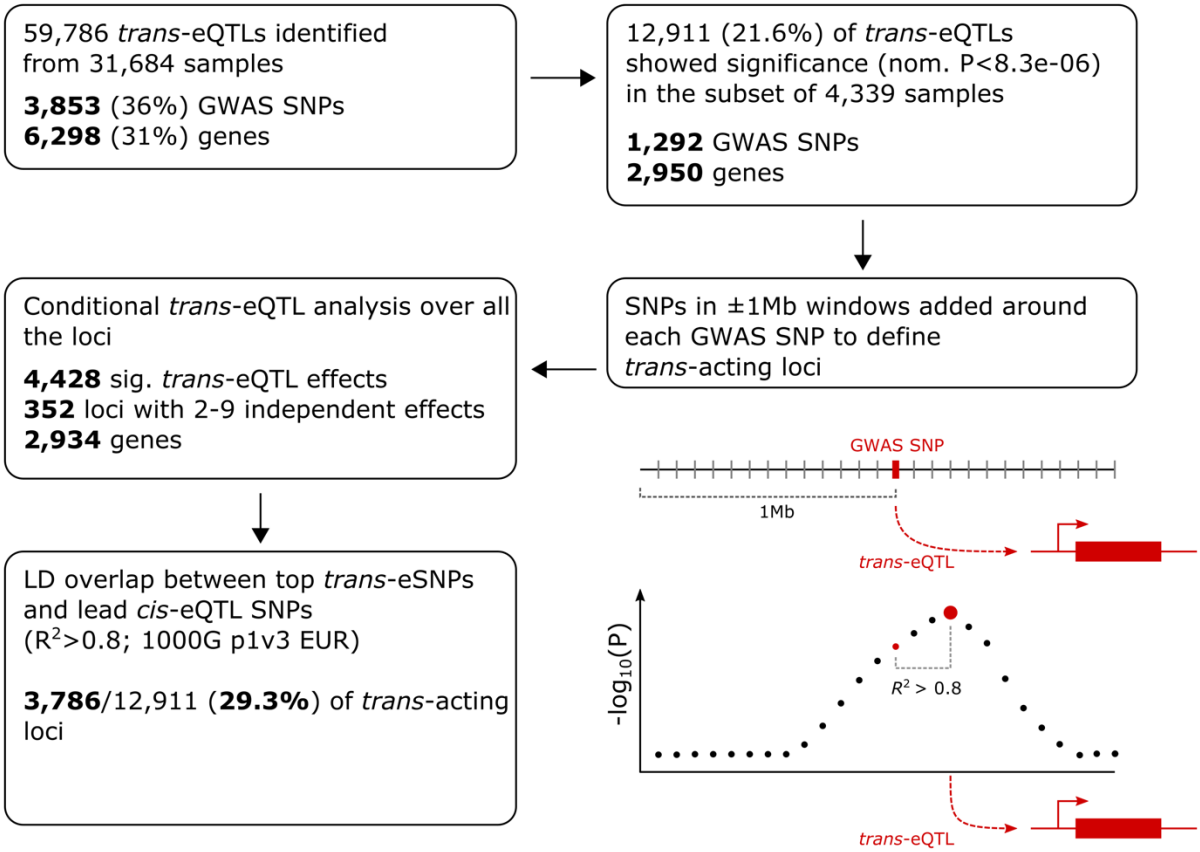
Supplementary Figure 16A. Comparison of the allelic direction of *cis*-eQTL lead SNPs between each individual dataset (y-axis) and the meta-analysis (x-axis) using Z-scores. Note, that for better visualization, the scale of the scales of y-axis differ from scales of x-axis. Black dots indicate *cis*-eQTLs with identical allelic direction compared to the meta-analysis, while red dots indicate opposite effects. *Cis*-eQTL directions per dataset are highly concordant with the meta-analysis.



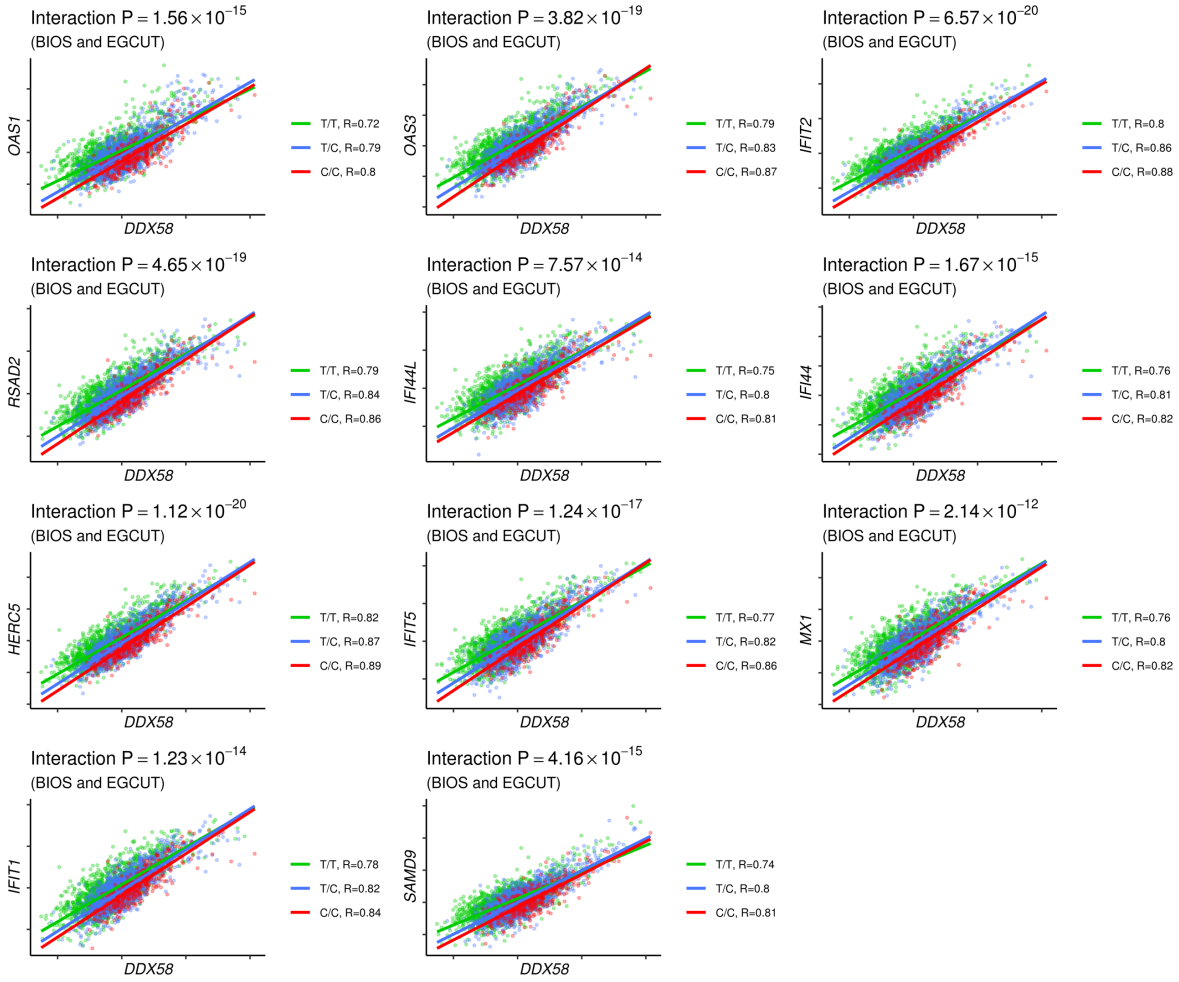
Supplementary Figure 16B. Comparison of the *trans*-eQTL allelic direction between each individual dataset (y-axis) and the meta-analysis (x-axis) using Z-scores. Note, that for better visualization, the scales of y-axis differ from scales of x-axis. Black dots indicate *trans*-eQTLs with identical allelic direction compared to the meta-analysis, while red dots indicate opposite effects. *Trans*-eQTL directions per dataset are highly concordant with the meta-analysis.



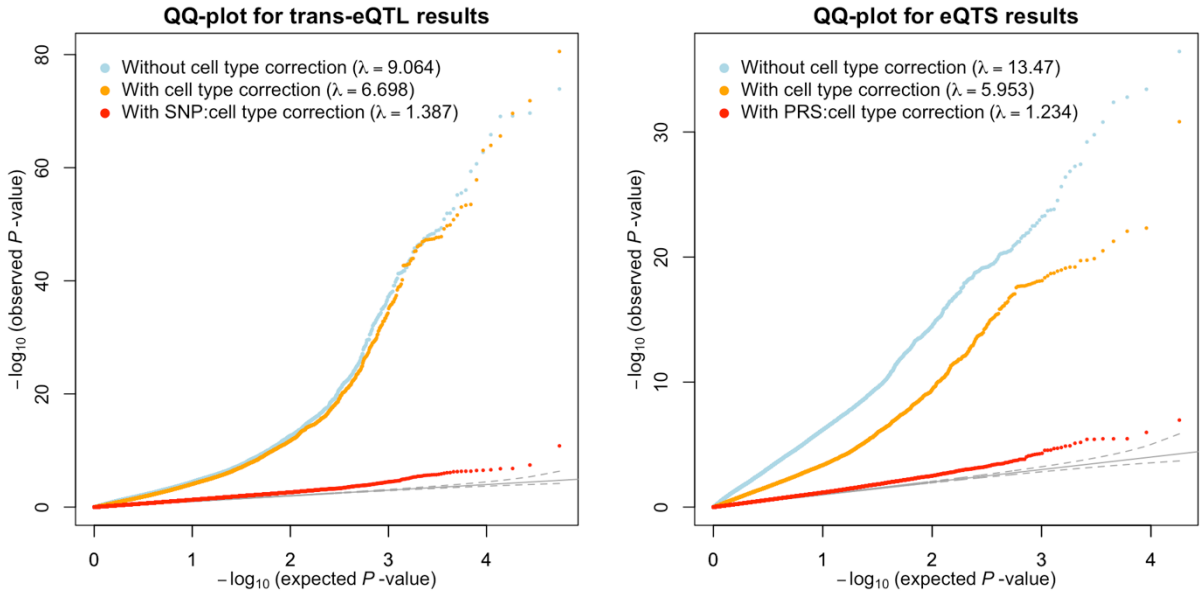
Supplementary Figure 16C. Comparison of the eQTS effect direction between each individual dataset (y-axis) and the meta-analysis (x-axis) using Z-scores. Note, that for better visualization, the scale of the scales of y-axis differ from scales of x-axis. Black dots indicate eQTSs with identical effect direction compared to the meta-analysis, while red dots indicate opposite effects. eQTS effect directions per dataset are highly concordant with the meta-analysis.



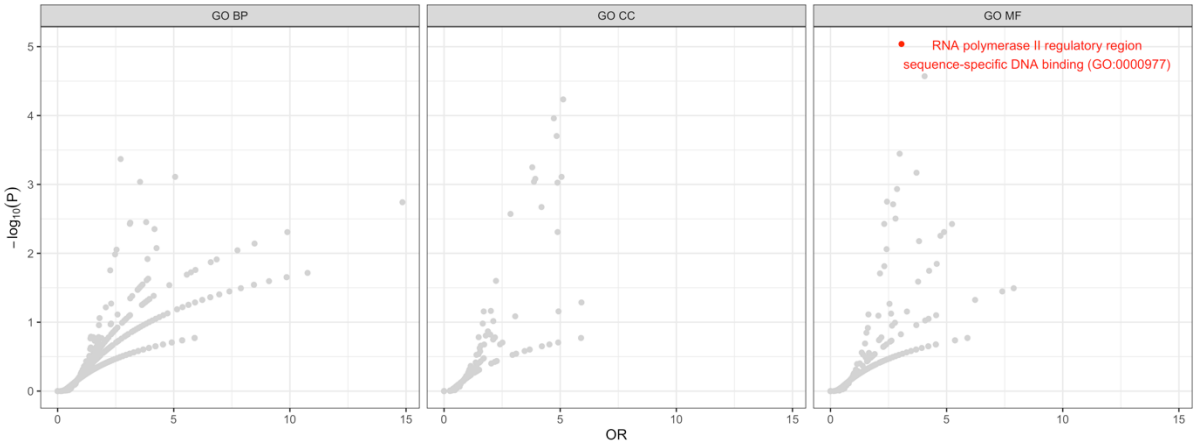
Supplementary Figure 17. Schematic of the conditional *trans*-eQTL analysis and LD-overlap with *cis*-eQTL SNPs. We analyzed the 59,786 *trans*-eQTLs identified in discovery meta-analysis, and limited ourselves to those having $P<8.3\times10^{-6}$ in the subset of 4,339 samples for which we had access to genotype and RNA-seq data. We then focused on regions of 1 Mb surrounding the identified *trans*-eQTL SNPs, and performed conditional analysis using all variants in those regions. Finally, we calculated LD between the independent lead variants showing strongest association in each locus and lead *cis*-eQTL SNPs from discovery analysis.



Supplementary Figure 18. *Trans*-eQTLs effects emerging from rs7045087 (9p21.1) are affected by *cis*-eQTL gene *DDX58*. For these analyses, we constructed *trans*-eQTL linear models and included *DDX58* into genotype × gene expression interaction term for each model. The two-sided P-value for genotype × *DDX58* interaction term is indicated for the each plot. A low P-value for this interaction term indicates that the three slopes (one for each genotype; red, blue and green) for the *trans*-eQTL differ depending on *DDX58* expression levels, suggesting that *DDX58* affects the outlined *trans*-eQTL effects. Interaction P-values on this plot were calculated based on the meta-analysis on all BIOS cohorts and EGCUT RNA-seq (combined N=4,339, all FDR<0.05). Scatterplots, slopes and corresponding Pearson correlation coefficients were calculated on BIOS cohorts only (combined N=3,831) because these RNA-seq data were generated and processed together.



Supplementary Figure 19. QQ-plots to investigate the effect of cell type composition on *trans*-eQTLs (left) and eQTS (right). We included only those discovery effects (FDR<0.05 in discovery meta-analysis) that were testable in up to 1,858 BIOS samples and SNPs which had per-cohort MAF>0.05. For both *trans*-eQTLs and eQTS, we evaluated three linear models (**Supplementary Methods**): a model without cell type correction (blue), a model including all 49 available cell metrics as covariates (orange), and a model including all 49 cell metrics and their respective SNP × cell-type or PGS × cell-type interaction terms as covariates (red). The y-axis shows the two-sided -$\log_{10}$(P-value) for the effect of the SNP or eQTS in each model. The x-axis shows the expected -$\log_{10}$(P-value) under a uniform null distribution. Lambda values indicate inflation over the null, and show decrease when interaction terms are added to the model, indicating that *trans*-eQTLs and eQTSs are likely cell type dependent.



Supplementary Figure 20. *Cis*-eQTL genes showing evidence of co-localization with *trans*-eQTLs (R^2>0.8, 1000G p1v3 EUR) were enriched by transcription factor process as defined by Gene ontology's Biological Process (GO BP), Cellular Compartment (GO CC) and Molecular Function (GO MF) categories. The x-axis shows the odds ratio and the y-axis shows the accompanying P-value ($-\log_{10}(P)$) from a one-sided Fisher's exact test and significant GO terms are outlined as red. In this analysis, the only effect reaching Benjamini-Hochberg FDR threshold of 0.05 had a P-value of 9.15×10^{-6}.

Discovery Cohorts

Illumina array cohorts

BEST

The BEST (Bangladesh Vitamin E and Selenium Trial) study is a randomized chemoprevention trial evaluating the long-term effects of vitamin E and selenium supplementation on non-melanoma skin cancer risk among 7,000 individuals with arsenic-related skin lesions living in seven sub-districts in Bangladesh (Argos et al., 2013). Participants included in this work are a subset of BEST participants for whom data is available on genome-wide single nucleotide polymorphisms (SNPs) and array-based expression. DNA was extracted from the whole blood using the QIAamp 96 DNA Blood Kit (cat # 51161; Qiagen, Valencia, USA). Concentration and quality of extracted DNA were assessed using Nanodrop 1000. Genotyping was conducted using Illumina HumanCytoSNP-12 v2.1 chips according to Illumina's protocol, and chips were read on the BeadArray Reader. Image data was processed in BeadStudio software to generate genotype calls. Quality control (QC) was conducted as described previously (Pierce et al., 2012; Pierce et al., 2013). We removed DNA samples with call rates <97%, gender mismatches, and technical duplicates. We removed SNPs with call rates <95% or HWE P-values<10^{-10}. The Michigan Imputation Server (Das et al., 2016) was used to conduct genotype imputation using 1,000 genomes reference haplotypes (1KG phase3 v5, mixed populations). Only high-quality imputed SNPs (imputation r2>0.3) with SNPs with Minor Allele Frequency (MAF) >0.05 were retained. RNA was extracted from PBMCs, preserved in buffer RLT, and stored at −86°C using RNeasy Micro Kit (cat# 74004) from Qiagen. Concentration and quality of RNA samples were assessed on Nanodrop 1000. cRNA synthesis was done from 250 ng of RNA using the Illumina TotalPrep 96 RNA Amplification kit, and 750 ng of cRNA was applied to the Illumina Human HT-12 v4 expression array. Individuals having <30% of probes with detection P-value <0.05 were excluded from the analysis. We also exclude 1st degree relatives by using GCTA software (–grm cut point of 0.3).

References

Argos, M., Rahman, M., Parvez, F., Dignam, J., Islam, T., Quasem, I., … Ahsan, H. (2013). Baseline comorbidities in a skin cancer prevention trial in Bangladesh. *European Journal of Clinical Investigation*, *43*(6), 579–588. http://doi.org/10.1111/eci.12085

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nature Genetics, 48(10), 1284–1287. http://doi.org/10.1038/ng.3656

Pierce, B. L., Kibriya, M. G., Tong, L., Jasmine, F., Argos, M., Roy, S., … Ahsan, H. (2012). Genome-Wide Association Study Identifies Chromosome 10q24.32 Variants Associated with Arsenic Metabolism and Toxicity Phenotypes in Bangladesh. PLoS Genetics, 8(2), e1002522. http://doi.org/10.1371/journal.pgen.1002522

Pierce, B. L., Tong, L., Argos, M., Gao, J., Farzana, J., Roy, S., … Ahsan, H. (2013). Arsenic metabolism efficiency has a causal role in arsenic toxicity: Mendelian randomization and gene-environment interaction. International Journal of Epidemiology, 42(6), 1862–71.

Acknowledgements
We would like to thank all BEST study participants and research staff. This work was supported by National Institutes of Health grants R21ES024834 (B.P. and M.A.), R01ES020506 (B.P.), R01ES023834 (B.P.), R35ES028379 (B.P.), R01 GM108711 (L.C.), and R01CA107431 (H.A.).

BSGS

The Brisbane Systems Genetics Study (BSGS) cohort was previously described in detail in (Powell et al., 2012; Powell et al., 2013). Briefly, BSGS is comprised of 862 individuals of Northern-European origin from 274 families consisting of either monozygotic or dizygotic twin pairs along with their siblings and parents. Expression levels for each individual were measured from whole blood using Illumina HT-12 v4.0 microarray chips. Whole genome SNP genotypes were generated using Illumina 610 Quad-Beadchips and, after QC, were imputed to the 1000 Genomes Release. The expression dataset is available from the GEO (Gene Expression Omnibus) public repository under the accession GSE 33321. Here, we selected only unrelated individuals, leaving 329 for analysis by the eQTLGen analysis plan for Illumina arrays.

References

Powell, J. E., Henders, A. K., McRae, A. F., Caracella, A., Smith, S., Wright, M. J., … Montgomery, G. W. (2012). The Brisbane Systems Genetics Study: Genetical Genomics Meets Complex Trait Genetics. *PLoS ONE*, *7*(4), e35430. http://doi.org/10.1371/journal.pone.0035430

Powell, J. E., Henders, A. K., McRae, A. F., Kim, J., Hemani, G., Martin, N. G., … Visscher, P. M. (2013). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. *PLoS Genetics*, *9*(5), e1003502. http://doi.org/10.1371/journal.pgen.1003502

Cardiology

Cardiology data contains 338 samples (Kim et al., 2014), and gene expression data have been deposited at the Gene Expression Omnibus archive under accession number (GEO:GSE49925). In this study, only 147 individuals in the discovery cohort were used. Peripheral blood samples were collected immediately prior to angiography and after overnight fasting, and stored in Paxgene tubes (QIAGEN, San Diego, CA, USA) at -80°C. Microarray analysis of transcript abundance was performed by hybridization of dye-labeled RNA to Illumina HT-12 v4 bead arrays. Hybridizations was performed by Expression Analysis (Durham, NC, USA). Whole genome genotypes for the discovery phase were determined by Illumina OmniQuad arrays at Expression Analysis (Durham, NC, USA), and was then imputed using Impute v2, using the 1000G reference phase1 v3 genotypes. After outlier detection and mixup correction, 134 individuals were included for the further analysis.

References

Kim, J., Ghasemzadeh, N., Eapen, D. J., Chung, N., Storey, J. D., Quyyumi, A. A., & Gibson, G. (2014). Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. *Genome Medicine*, *6*(5), 40. http://doi.org/10.1186/gm560

CHDWB

The Center for Health Discovery and Well Being (CHDWB) cohort (Preininger et al., 2013, Wingo et al., 2015) is comprised of 465 samples collected in Atlanta, Georgia, USA. Whole peripheral blood RNA samples were collected using Tempus Blood RNA Tubes (Life Technologies, NY, USA), and RNA was extracted using Tempus Spin RNA Isolation Kit (Life Technologies, NY, USA). Quality was measured by NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Whole-Genome gene-

expression levels were available for 408 individuals and were obtained by Illumina Human HT-12 v4 arrays (Illumina Inc, San Diego, US) according manufacturers protocols. Genotyping was performed using HumanCoreExome and HumanOmniExpress, and was imputed using Impute v2, using the 1000G reference phase1 v3 genotypes. After outlier detection and mix-up correction, 384 samples were left for eQTL detection.

References

Preininger, M., Arafat, D., Kim, J., Nath, A. P., Idaghdour, Y., Brigham, K. L., & Gibson, G. (2013). Blood-Informative Transcripts Define Nine Common Axes of Peripheral Blood Gene Expression. *PLoS Genetics*, *9*(3), e1003362. https://doi.org/10.1371/journal.pgen.1003362

Wingo, A. P., & Gibson, G. (2015). Blood gene expression profiles suggest altered immune function associated with symptoms of generalized anxiety disorder. *Brain, Behavior, and Immunity*, *43*, 184–191. https://doi.org/10.1016/j.bbi.2014.09.016

Morocco

The Morocco dataset (Idaghdour et al., 2010) is comprised of 194 individuals representing two ethnicities (Arabs and Amazighs) from three geographic locations (Agadir, Boutroch and Ighrem) and two lifestyles (Urban and Rural). Gene expression data from this study have been deposited under series GSE17065. Peripheral blood samples were collected over the course of 6 d during June and July 2008 using LeukoLock™ system (https://www.thermofisher.com/order/catalog/product/AM1923#/AM1923). The same collection protocol was followed for all samples to minimize heterogeneity due to technical reasons. Total RNA extraction and cDNA and cRNA synthesis were performed with an Illumina TotalPrep RNA Amplification kit (Ambion) in accordance with the manufacturer's instructions. Total RNA samples were checked for quality with an RNA 6000 Nano LabChip kit and 2100 Bioanalyzer (Agilent). RNA from each was hybridized to an Illumina HT-12 v3 array. Genotype was assayed with Infinium Human 610-Quad beadchips (Illumina) by following standard procedures, also at the Duke University IGSP. The beadchips were imaged by using a BeadArray Reader (Illumina), and genotype calls were extracted with the Genotyping Module in BeadStudio software, and was then imputed using Impute v2, using the 1000G reference phase1 v3 genotypes. After outlier detection and mix-up correction, a total of 175 individuals were retained.

References

Idaghdour, Y., Czika, W., Shianna, K. V., Lee, S. H., Visscher, P. M., Martin, H. C., … Gibson, G. (2010). Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nature Genetics*, *42*(1), 62–67. https://doi.org/10.1038/ng.495

EGCUT

The Illumina array gene expression data from the Estonian Genome Center, University of Tartu (EGCUT, Leitsalu et al., 2015) biobank of 53,000 samples consists of two different cohorts: the Estonian Gene Expression Cohort (EGCUT1) and the Center for Translational Genomics cohort (EGCUT2). EGCUT Illumina array analyses were approved by Ethics Review Committee of Human Research of the University of Tartu, Estonia (permission no 234/T-12).

EGCUT1

EGCUT1 cohort is composed of 2,658 individuals obtained from the Estonian Genome Center, University of Tartu (EGCUT) cohort. Genotyping was performed using Human370CNV BeadChips (Illumina), and imputed using the 1000 Genomes project reference by IMPUTE v2 (Phase III, March 2012 release). SNP QC was done on the basis of call rate (≤97%) across samples and deviation from HWE ($\leq 1 \times 10^{-6}$). Before imputation, non-autosomal SNPs, SNPs with minor allele frequency < 1% and palindromic SNPs were removed. Gene expression levels were measured in whole blood using Illumina HT12v3 microarray chips. After sample QC and mix-up correction, 818 samples were included in the eQTLGen meta-analyses. The expression dataset is available at GEO public repository under the accession GSE48348.

EGCUT2

EGCUT2 cohort is composed of 1,000 individuals who have been re-contacted for a second time-point sample. Of these, 96 individuals have gene expression levels measured in whole blood. Genotyping was performed using HumanOmniExpress BeadChips (Illumina), and imputed using the 1000 Genomes project reference by IMPUTE v2. RNA from whole blood was purified using the MagMAX™ for Stabilized Blood Tubes RNA Isolation Kit. RNA was concentrated using the Heraeus vacuum centrifugation system without heating. RNA was labeled and amplified using the TargetAmp-Nano Labeling Kit for Illumina Expression BeadChip (Epicentre Biotechnologies) with SuperScript III Reverse Transcriptase (Life Technologies), followed by purification with the RNeasy MinElute Cleanup Kit (Qiagen). RNA quality was assessed after extraction and after labelling using an Agilent 2100 Bioanalyzer and Agilent RNA 6000 Nano Kit (all from Agilent Technologies). Labeled RNA was hybridized to the HumanHT-12 v4 Expression BeadChip (Illumina) according to the manufacturer's instructions. After sample QC and mix-up correction, 19 samples were excluded and 77 samples remained. The expression dataset is available at GEO public repository under the accession GSE78840.

References

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., … Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. http://doi.org/10.1093/ije/dyt268

Acknowledgements

EGCUT analyses were funded by EU H2020 grant 692145, Estonian Research Council Grant IUT20-60, IUT24-6, and European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012GENTRANSMED.

This work was carried out in the High Performance Computing Center of University of Tartu.

DILGOM

The DIetary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) cohort (Inouye et al., 2010) is composed of 518 individuals obtained from the National FINRISK Study, The National Institute for Health and Welfare, Helsinki, Finland. Genotyping was performed using HumanHap 610K BeadChips (Illumina), and imputed using the 1000 Genomes project reference by IMPUTE v2 (Phase III, March 2012 release). SNP QC was done on the basis of call rate (≤98%) across samples and deviation from HWE ($\leq 1 \times 10^{-6}$). Before imputation, non-autosomal SNPs, SNPs with minor allele frequency < 1% and palindromic SNPs were removed. Gene expression levels were measured in whole blood using Illumina HT12v3 microarray chips. After sample QC and mix-up correction, 498 samples were included to eQTLGen meta-analyses.

References

Inouye, M., Silander, K., Hamalainen, E., Salomaa, V., Harald, K., Jousilahti, P., … Peltonen, L. (2010). An Immune Response Network Associated with Blood Lipid Levels. *PLoS Genetics*, *6*(9), e1001113. https://doi.org/10.1371/journal.pgen.1001113

Fehrmann

The Fehrmann datasets consist of whole blood samples from United Kingdom and Netherlands (Dubois et al., 2010; Fehrmann et al., 2011). This dataset consists of blood samples from patients (overview: Fehrmann et al., 2011) and healthy controls. Samples were genotyped with Illumina HumanHap300, HumanHap370 or 610 Quad platforms. Genotypes were imputed by Impute v2 (Howie et al., 2009) using the GIANT 1000G p1v3 integrated call set for all ancestries as a reference (The 1000 Genomes Project Consortium, 2010). Gene expression levels were measured by Illumina HT-12 v3 and Illumina HumanRef-8 v2.0 arrays. This expression dataset is available at GEO (Gene Expression Omnibus) repository, accession numbers GSE20142 and GSE20332. After data processing and QC, 1,227 samples from HT-12 v3 dataset and 214 from HumanRef-8 v2.0 dataset were included to eQTLGen meta-analyses. All samples were collected after informed consent and approved by local ethical review boards.

References

Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., … van Heel, D. A. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, *42*(4), 295–302. http://doi.org/10.1038/ng.543

Fehrmann, R. S. N., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., … Franke, L. (2011). Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genetics*, *7*(8), e1002197. http://doi.org/10.1371/journal.pgen.1002197

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5*(6), e1000529. http://doi.org/10.1371/journal.pgen.1000529

The 1000 Genomes Project Consortium, Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. http://doi.org/10.1038/nature09534

HVH

The Heart and Vascular Health (HVH) study (Heckbert et al., 2009; Psaty et al., 1995; Smith et al., 2004) constitutes a group of population-based case control studies of myocardial infarction (MI), stroke, venous thromboembolism (VTE), and atrial fibrillation. Study participants were 30-79 year old members of Group Health, a large integrated health care organization in Washington State. Cases were identified from hospital discharge diagnosis codes and subsequently validated by medical record review. Cases shared a common control group that was a random sample of Group Health members, frequency-matched to MI cases on age (within decade), sex, treated hypertension, and calendar year of identification. The HVH study started in 1987 and blood specimens have been collected since 1995. Study eligibility, participant characteristics and risk

factor information were collected by medical record review and telephone interview. In addition, surviving cases and controls who agreed to participate had blood drawn.

Since 2003, whole blood has been collected in PAXGene tubes for mRNA expression studies. Participants of the current study are those for whom expression profiling was done as part of several gene expression pilot studies conducted among HVH controls to investigate incident cardiovascular disease, hormone therapy, medications, diabetes and atrial fibrillation. The Group Health human subjects review committee approved the study and all participants provided written informed consent.

Total RNA was extracted using PAXGene Blood RNA Kit and RNase-Free DNase Set (QIAGEN Inc., Valencia, CA) at the Fred Hutchinson Cancer Research Center, Seattle, WA. RNA integrity and quality was assessed using Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA). Illumina® TotalPrep™-96 RNA Amplification Kit (Life Technologies Corp., Carlsbad, CA) was used for RNA amplification and labeling using manufacturer's instructions. Labeled cRNAs were hybridized onto Illumina HumanHT-12v3 and Illumina HumanHT-12 v4 Expression Beadchip (Illumina, San Diego, CA) arrays, according to manufacturer's protocols. The images of the array chips were captured using an Illumina Beadarray scanner and scanned array images were imported into Illumina's GenomeStudio Gene Expression Module. RNA QC and microarray expression profiling experiments were conducted at the laboratory of Dr. Jerome Rotter. The expression data is available at GEO (Gene Expression Omnibus) public repository under the accession GSE47729.

Genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina 370CNV BeadChip system. Genotypes were called using the Illumina BeadStudio software. Samples were excluded from analysis for sex mismatch or call rate < 95%. The following exclusions were applied to identify a final set of 301,321 autosomal SNPs: call rate < 97%, HWE P-value < 10-5, > 2 duplicate errors or Mendelian inconsistencies (for reference CEPH trios), heterozygote frequency = 0, SNP not found in HapMap and inconsistencies across genotyping batches. The genotypes retained after QC were pre-phased using MaCH. The phased genotypes were imputed into a reference panel of 1092 individual of multiple ethnicities from the Phase1 version 3 haplotypes of Thousand Genomes project using minimac (release stamp 2012-11-16). Imputation of the X chromosome was limited to the non pseudo-autosomal region and was imputed separately by sex. Samples profiled by HT-12 v3 (N=40) and HT-12 v4 (N=59) arrays were analyzed separately.

References

Heckbert, S. R., Wiggins, K. L., Glazer, N. L., Dublin, S., Psaty, B. M., Smith, N. L., … Lumley, T. (2009). Antihypertensive Treatment With ACE Inhibitors or -Blockers and Risk of Incident Atrial Fibrillation in a General Hypertensive Population. *American Journal of Hypertension*, *22*(5), 538–544. http://doi.org/10.1038/ajh.2009.33

Psaty, B. M., Heckbert, S. R., Koepsell, T. D., Siscovick, D. S., Raghunathan, T. E., Weiss, N. S., … Wahl, P. W. (1995). The risk of myocardial infarction associated with antihypertensive drug therapies. *JAMA*, *274*(8), 620–5.

Smith, N. L., Heckbert, S. R., Lemaitre, R. N., Reiner, A. P., Lumley, T., Weiss, N. S., … Psaty, B. M. (2004). Esterified Estrogens and Conjugated Equine Estrogens and the Risk of Venous Thrombosis. *JAMA*, *292*(13), 1581. http://doi.org/10.1001/jama.292.13.1581

Acknowledgments

HVH was supported in part by grants R01 HL085251 and R01 HL073410 from the National Heart, Lung, and Blood Institute (NHLBI). The provision of genomic data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and National

Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHLBI or the National Institutes of Health, USA. Dr. Psaty serves on a DSMB for a clinical trial of a device funded by Zoll LifeCor and on the Steering Committee for the Yale Open Data Access Project funded by Johnson & Johnson. Both activities are unrelated to this work.

InCHIANTI

The Invecchiare in Chianti, ageing in the Chianti area (InCHIANTI) study (http://www.inchiantistudy.net) (Ferrucci et al., 2000) is a population-based, prospective study of human ageing in the Tuscany region of Italy. 1,455 participants were enrolled at baseline (1998-2000), with follow-up waves every 3 years. Extensive interviews, questionnaires, medical examinations, physical tests and blood samples were taken at every wave. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board in Italy, and participants gave informed consent to participate.

At wave 4 (year 9, 2008/9), peripheral blood specimens were collected from 712 individuals using the PAXGene technology to preserve levels of mRNA transcripts as they were at the point of collection (Debey-Pascher et al., 2009). RNA was extracted from peripheral blood samples using the PAXGene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer's instructions. RNA was biotinylated and amplified using the Illumina® TotalPrep (Broer et al., 2014) -96 RNA Amplification Kit and directly hybridized with HumanHT-12 v3 Expression BeadChips that include 48,803 probes. Image data was collected on an Illumina iScan and analyzed using the Illumina and Beadstudio software (Illumina, San Diego, California, USA) as previously described (Gibbs et al., 2010). All microarray experiments and analyses complied with MIAME guidelines. Genotyping was carried out by Illumina 550K array and imputation was performed by MACH, using 1000G phase 1 v3 reference panel.

		Inclusion criteria					Imputation inclusion criteria	
Genotype Platform	Calling algorithm	MAF	Call rate	HWE p-value	N of SNPs that met QC criteria	Imputation software	MAF	
Illumina 550K	Beadstudio	≥1%	≥99%	>10-6	514,027	MACH	≥1%	R2-hat≥0.30

The total number of InCHIANTI samples with good quality whole-genome expression data equals 698, 695 of which also have cell-count data. After preprocessing and QC, 609 samples were included to eQTLGen study.

References

Broer, L., Buchman, A. S., Deelen, J., Evans, D. S., Faul, J. D., Lunetta, K. L., … Murabito, J. M. (2015). GWAS of Longevity in CHARGE Consortium Confirms APOE and FOXO3 Candidacy. *The Journals of Gerontology: Series A*, *70*(1), 110–118. http://doi.org/10.1093/gerona/glu166

Debey-Pascher, S., Eggle, D., & Schultze, J. L. (2009). RNA Stabilization of Peripheral Blood and Profiling by Bead Chip Analysis. In *Methods in molecular biology (Clifton, N.J.)* (Vol. 496, pp. 175–210). http://doi.org/10.1007/978-1-59745-553-4_13

Ferrucci, L., Bandinelli, S., Benvenuti, E., Di Iorio, A., Macchi, C., Harris, T. B., & Guralnik, J. M. (2000). Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *Journal of the American Geriatrics Society*, *48*(12), 1618–25.

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., … Singleton, A. B. (2010). Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *PLoS Genetics*, *6*(5), e1000952. http://doi.org/10.1371/journal.pgen.1000952

Acknowledgements

This study was supported in part by the Intramural Research Program, National Institute on Ageing. UK-based work was generously supported by a Wellcome Trust Institutional Strategic Support Award (WT097835MF), plus internal University of Exeter Medical School funding. This work has made use of the resources provided by the University of Exeter Science Strategy and resulting Systems Biology initiative.

KORA F4

The Cooperative Health Research in the Region of Augsburg (KORA F4) is a follow-up survey (2006-2008) of the population-based KORA S4 survey that was conducted in the region of Augsburg in Southern Germany in 1999-2001. The expression analysis in this study was based on whole blood samples of the KORA F4 participants aged 62 to 81 years (Rathmann et al., 2009). RNA was isolated from whole blood using PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany). Purity and integrity of the RNA was analyzed using the Agilent Bioanalyzer with the 6000 Nano LabChip reagent set (Agilent Technologies, Germany). RNA was reverse transcribed with TotalPrep-96 RNA Amp Kit (Ambion, Germany) and hybridized to the Illumina HumanHT-12 v3 Expression BeadChip (Schurmann et al., 2012). The samples were genotyped on the Affymetrix 6.0 GeneChip array and imputed with IMPUTE (v1.0.15) using 1000 Genomes phase 1, version 3 as reference population for calling and imputation. All together there were 952 samples with gene expression and genotype data available for analysis. The expression dataset is available at ArrayExpress public repository under the accession E-MTAB-1708.

References

Rathmann, W., Strassburger, K., Heier, M., Holle, R., Thorand, B., Giani, G., & Meisinger, C. (2009). Incidence of Type 2 diabetes in the elderly German population and the effect of clinical and lifestyle risk factors: KORA S4/F4 cohort study. *Diabetic Medicine*, *26*(12), 1212–1219. http://doi.org/10.1111/j.1464-5491.2009.02863.x

Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., … Ziegler, A. (2012). Analyzing Illumina Gene Expression Microarray Data from Different Tissues: Methodological Aspects of Data Analysis in the MetaXpress Consortium. *PLoS ONE*, *7*(12), e50938. http://doi.org/10.1371/journal.pone.0050938

Acknowledgements

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was

supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

LIFE-Adult

LIFE-Adult is a population-based cohort study that recruited about 10,000 randomly selected inhabitants of the city of Leipzig, Saxony, Germany. Further information can be found elsewhere (Loeffler et al., 2015).

Gene expression

Whole blood was collected in Tempus Blood RNA Tubes (Life Technologies) and relocated to -80°C before further processing. Isolated RNA was processed and hybridized to Illumina HT-12 v4 Expression BeadChips (Illumina, San Diego, CA, USA) and measured on the Illumina HiScan.

Raw data of all 47,322 probes was extracted by Illumina GenomeStudio in all 3,489 initially included individuals. Three criteria were used to remove samples of low quality (processed within R / Bioconductor): First, the number of gene-expression probes detected in a sample was required to be within ± 3 interquartile ranges (IQR) from the median, leading to the exclusion of 0.4% of the initially included samples. Second, as described in (Du et al., 2008), log-transformed and quantile-normalized Euclidean distances of expression values had to be within 4 x IQR of the median, leading to the exclusion of 0.2% of the samples. Third, the Mahalanobis distance of several quality characteristics of each sample (log-transformed and quantile-normalized signal of perfect-match and miss-match control probes, control probes present at different concentrations, mean of negative control probes, mean of house-keeping genes, number of expressed genes, mean signal strength of biotin-control-probes and ERCC-spike-in probes (Jiang et al., 2011) had to be within median + 3 x IQR, leading to the exclusion of an additional 2.3% of the samples. Hence, valid expression data was available in a total of 3,388 individuals.

Genotyping

Genomic DNA was extracted from peripheral blood leukocytes applying an automated protocol on the Autopure LS instrument (Qiagen, Hilden, Germany) as recommended by the manufacturer. Chip-genotyping was done applying Axiom Genome-Wide CEU 1 Array Plate (Affymetrix, Inc., Santa Clara, California, USA) technology including 587,352 Single Nucleotide Polymorphisms (SNPs) according the manufacturer's instructions. Sample quality filtering removed all individuals with dish-QC < 0.82, call rate < 0.97, reported vs. genotype-wise computed sex mismatch and cryptic relatedness. Using about 200,000 high-quality SNPs (call rate > .998), PCA was performed using EIGENSOFT 3.0. Outliers according to the standard-cutoff 6SD were removed, leaving 4985 individuals for further analysis. SNP quality filtering removed SNPs with call rate < 0.97, Hardy-Weinberg P-value ≤ 1E-6, plate association p-value ≤ 1E-7 and inappropriate cluster-plot quality metrics (Fisher's Linear Discriminant < 3.6, heterozygous cluster strength offset < -0.1 or invalid homozygote ratio offset). A total of 538,181 mapped SNPs were included in imputation using 1000G reference phase 1, release V3 of CEU as reference (hg19, dbSNP 135). Data were imputed by first pre-phasing using SHAPEIT (version v2.r778) with standard settings for European populations followed by imputation with IMPUTE2 (version 2.3.0), resulting in 39,300,191 imputed SNPs. For post-imputation QC, SNPs with minor allele frequency<1% or info-score<0.5 were removed.

For eQTL analysis, individuals with both genotyping and expression data were filtered. After validating that no mix-up of gene expression and genetic data was present, a total of 1,978 individuals were included into interim *cis*-eQTL analysis and 2,456 individuals into final *cis*-eQTL, *trans*-eQTL and eQTS analyses.

References

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547–1548. http://doi.org/10.1093/bioinformatics/btn224

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., … Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, *21*(9), 1543–1551. http://doi.org/10.1101/gr.121095.111

Loeffler, M., Engel, C., Ahnert, P., Alfermann, D., Arelin, K., Baber, R., … Thiery, J. (2015). The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC Public Health*, *15*(1), 691. http://doi.org/10.1186/s12889-015-1983-z

LIFE-Heart

LIFE-Heart is a cohort study of patients with suspected or confirmed coronary artery disease collected at the Heart-Center of Leipzig. Details of the study can be found elsewhere (Scholz et al., 2020).

Gene expression

PBMC-based gene-expression was measured by Illumina HumanHT-12 v4 Expression BeadChip. Pre-processing was performed analogously to LIFE-Adult as described previously (Kirsten et al., 2015). Briefly, PBMC isolation was performed using Cell Preparation Tubes (CPT, Becton Dickinson), total RNA was extracted using TRIzol reagent (Invitrogen) and 500 ng RNA per sample were ethanol precipitated with GlycoBlue (Invitrogen) as carrier and hybridized to Illumina HT-12 v4 Expression BeadChips (Illumina, San Diego, CA, USA). Raw data of all 47,231 gene-expression probes available was extracted by Illumina GenomeStudio without additional background correction. Three criteria were used to remove samples of low quality (processed within R / Bioconductor): First, the number of detected gene-expression probes of a sample was required to be within ± 3 interquartile ranges (IQR) of the median, leading to the exclusion of 2.7% of the 4,509 initially included samples. Second, the Mahalanobis distance of several quality characteristics of each sample (log-transformed and quantile-normalized signal of biotin-control-probes, signal of low-concentration control probes, signal of medium-concentration control probes, signal of mismatch control probes, signal of negative control probes and signal of perfect-match control probes) (Cohen et al., 2007) had to be within median + 3 x IQR, leading to the exclusion of an additional 0.02% of the samples. Third, log-transformed and quantile-normalized Euclidean distances of expression values as described (Du et al., 2008) had to be within 4 × IQR from the median, leading to the exclusion of another 0.4% of the samples. Overall, of the 4,509 samples assayed, 141 samples were excluded for quality reasons.

Genotyping

Genotyping was performed using the Affymetrix Axiom Technology with custom option (Axiom-CADLIFE). Genotype calling was performed with Affymetrix Power Tools version 1.12. Sample QC comprised call rate (>97%), hetero- or homozygosity excess (outliers of mean squared differences of observed and expected genotypes), sex-mismatch, cryptic relatedness and outliers of PCA (6SD criterion of Eigenstrat software (Price et al., 2006)). Prior to imputation, low quality SNPs defined by low call-rate (<90% plate-wise call rate corresponding to <94.2% overall call-rate), deviation from HWE (P-value $<10^{-6}$) or plate-association (P-value $<10^{-7}$) were filtered. Individuals were imputed at the 1000Genomes reference phase 1, release 3 (http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html) using SHAPEIT v2 and IMPUTE 2.3.0. For post-imputation QC, SNPs with minor allele frequency<1% or info-score<0.5 were removed.

For eQTL analysis, individuals with both genotyping and expression data were filtered. After validating that no mix-up of gene expression and genetic data is present, a total of 2,106 individuals were included into interim *cis*-eQTL analysis and 4,285 individuals into final *cis*-eQTL, *trans*-eQTL and eQTS analyses.

References

Cohen Freue, G. V., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., … Ng, R. T. (2007). MDQC: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, *23*(23), 3162–3169. http://doi.org/10.1093/bioinformatics/btm487

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547–1548. http://doi.org/10.1093/bioinformatics/btn224

Kirsten, H., Al-Hasani, H., Holdt, L., Gross, A., Beutner, F., Krohn, K., … Scholz, M. (2015). Dissecting the genetics of the human transcriptome identifies novel trait-related *trans* -eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Human Molecular Genetics*, *24*(16), 4746–4763. http://doi.org/10.1093/hmg/ddv194

Scholz, M., Henger, S., Beutner, F., Teren, A., Baber, R., Willenberg, A., Ceglarek, U., Pott, J., Burkhardt, R., & Thiery, J. (2020). Cohort Profile: The Leipzig Research Center for Civilization Diseases-Heart study (LIFE-Heart). In *International Journal of Epidemiology* (Vol. 49, Issue 5, pp. 1439-1440H). Oxford University Press. https://doi.org/10.1093/ije/dyaa075

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. http://doi.org/10.1038/ng1847

Acknowledgments

We thank all study participants, volunteers and study personnel who made this work possible. We thank Lesca Holdt very much for helping to create the LIFE-Heart gene-expression data, and thank Kay Olischer and Annegret Unger very much for technical assistance. We thank Kerstin Wirkner very much for running the LIFE study center, and thank Sylvia Henger very much for LIFE-Adult and LIFE-Heart data quality control.

LIFE – Leipzig Research Centre for Civilization Diseases is an organizational unit affiliated to the Medical Faculty of the University of Leipzig. LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by funds of the Free State of Saxony within the framework of the excellence initiative (project numbers 713-241202, 713-241202, 14505/2470, 14575/2470).

Rotterdam Study

The Rotterdam Study (Hofman et al., 2015) is a single-center, prospective population-based cohort study conducted in Rotterdam, the Netherlands. Subjects were included in different phases from the start of the study in 1998, with a total of 14,926 men and women aged 45 years and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of risk factors for chronic diseases to contribute to better prevention and treatment of such diseases in the elderly.

Whole-blood was collected in PAXGene tubes (Becton Dickinson) and total RNA was isolated using PAXGene Blood RNA kits (Qiagen). To ensure constant high-quality of RNA preparation,

all RNA samples were analyzed using the Labchip GX (Calliper) according to manufacturer's instructions. Samples with an RNA Quality Score > 7 were amplified and labelled (Ambion TotalPrep RNA) and hybridized to the Illumina HumanHT-12 v4 Expression Beadchips (Illumina) as described by the manufacturer's protocol. Processing of the Rotterdam Study RNA samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Center Rotterdam. For genotyping, whole blood was also collected in EDTA tubes and DNA was isolated using a manual salting-out protocol. Genotyping for this sample subset was performed on the Illumina 610K quad beadchip array (Illumina) according to manufacturer's specifications. SNP genotype data was then imputed to a combined 1000 Genomes + UK10K imputation panel using IMPUTE2 software.

References

Hofman, A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., … Vernooij, M. W. (2015). The Rotterdam Study: 2016 objectives and design update. *European Journal of Epidemiology*, *30*(8), 661–708. http://doi.org/10.1007/s10654-015-0082-x

SHIP-Trend

The Study of Health in Pomerania (SHIP) is a population-based survey consisting of two independent cohorts in the North-East of Germany, SHIP and SHIP-Trend. The study design and sampling methods were described previously (Volzke et al., 2011). For this eQTL analysis, a subset of the SHIP-Trend cohort (N=986) with gene expression levels measured was used. Serum aliquots of the SHIP-Trend probands were prepared for immediate analysis and for storage at -80 °C in the Integrated Research Biobank (Liconic, Liechtenstein) and genotyped using the Illumina HumanOmni2.5 Quad arrays. Samples with a call rate < 94%, with reported vs. genotyped sex-mismatch, and duplicate samples (by estimated IBD) were excluded. Genotypes were imputed to 1000 Genomes v3 using IMPUTE v2.2.2. Prior to imputation, monomorphic SNPs, SNPs with a call rate ≤ 90%, and SNPs out of Hardy-Weinberg-Equilibrium ($p \leq 0.0001$) were excluded. Blood sample collection as well as RNA preparation were described in detail elsewhere (Schurmann et al. 2012). Briefly, RNA was prepared from whole blood under fasting conditions in PAXgene tubes (BD) using the PAXgene Blood miRNA Kit (Qiagen, Hilden, Germany) on a QIAcube according to the protocols provided by the manufacturer (Qiagen). RNA was amplified (Ambion TotalPrep RNA), and hybridized to the Illumina whole-genome Expression BeadChips (HT-12v3). The SHIP-Trend expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 36382. After sample mix-up correction and outlier removal based on genetic principal components (<4 SD from mean of PC1) both imputed genotypes and whole-blood gene expression data were available for a total of 955 SHIP-TREND samples. The medical ethics committee of the University of Greifswald approved the study protocol, and oral and written informed consents were obtained from each of the study participants.

References

Volzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., … Hoffmann, W. (2011). Cohort Profile: The Study of Health in Pomerania. *International Journal of Epidemiology*, *40*(2), 294–307. http://doi.org/10.1093/ije/dyp394

Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., … Ziegler, A. (2012). Analyzing Illumina Gene Expression Microarray Data from Different Tissues: Methodological Aspects of Data Analysis in the MetaXpress Consortium. *PLoS ONE*, *7*(12), e50938. http://doi.org/10.1371/journal.pone.0050938

Acknowledgements

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research (grant 03IS2061A).

SSIC

The Singapore Systems Immunology Cohort (SSIC) is a cross-sectional cohort collected by the Singapore Immunology Network (SIgN) at the Agency for Science, Technology and Research (A*STAR) along with National University of Singapore. The SSIC consists of ethnic Chinese participants recruited in Singapore for whom we have collected multiple high throughput data such as genomics, transcriptomics and metabolomics. The volunteer recruitment and sample collection have been described previously (Andiappan et al., 2014; Puan et al., 2017). For this eQTL analysis, whole blood gene expression data from the Singapore Chinese cohort was determined from the mRNA isolated from blood. Each blood sample was processed using the Tempus Spin Column Blood RNA isolation kit (Applied Biosystems) and the MagMAX Blood RNA extraction kit (Life Technologies). After which, globin mRNA reduction was conducted on a fraction of the extracted total RNA. The extracted RNA was then hybridized onto Illumina HumanHT-12 whole-genome gene expression chips. To avoid batch effects, the RNA samples were randomly placed onto Illumina HumanHT-12 arrays such that each chip contains a number of samples from the various RNA extraction batches. Additionally, whole genome genotyping was done on these individuals using the Human Illumina Omni 5M chip. Samples were then checked for any pair of samples identified as first-degree relatives, and, if found, these were removed. SNPs that were monomorphic in the population and those that failed a call rate of 95% were also removed. The genotyping data for 303 individuals was determined for 4,286,238 SNPs and taken forward for further statistical analysis.

References

Andiappan, A. K., Puan, K. J., Lee, B., Nardin, A., Poidinger, M., Connolly, J., … Rotzschke, O. (2014). Allergic airway diseases in a tropical urban environment are driven by dominant mono-specific sensitization against house dust mites. *Allergy*, *69*(4), 501–509. http://doi.org/10.1111/all.12364

Puan, K. J., Andiappan, A. K., Lee, B., Kumar, D., Lai, T. S., Yeo, G., … Rötzschke, O. (2017). Systematic characterization of basophil anergy. *Allergy*, *72*(3), 373–384. http://doi.org/10.1111/all.12952

Acknowledgements

The authors would like to thank all the volunteers and their family members who participated in this study. We would also like to thank Ramani Anantharaman, Parate Pallavi Nilkanth, Bani Kaur Suri, Sri Anusha Matta, and other members of the functional genomics laboratory at National University of Singapore for helping in sample collection. This study was supported by grants from the Singapore Immunology Network (SIgN-06-006, SIgN-08-020 and SIgN-10-029); the National Medical Research Council (NMRC/1150/2008), Singapore; the Biomedical Research Council, Singapore; SIgN core funding from the Agency for Science, Technology and Research (A*STAR); and the National University of Singapore for the Graduate Research Scholarship for students involved in the study.

Sorbs

The Sorbs are a population of Slavonic origin living in ethnic isolation among the Germanic majority in Eastern Saxony for about 1100 years. A convenience sample of this population was collected including unrelated subjects and families. Details of the population can be found elsewhere (Tonjes et al., 2009; Gross et al., 2011).

Gene expression

PBMCs were extracted from blood samples collected in VACUTAINER CPT (Cell Preparation Tubes) containing sodium heparin as the anti-coagulant according to the manufacturer's protocol (BD, Franklin Lakes, NJ). RNA from PBMCs was extracted using the TRIzol protocol (Thermo Fisher Scientific). DNase I digestion of the RNA samples was performed with subsequent RNA clean-up using the RNeasy MinElute Cleanup Kit (Qiagen, Hilden, Germany). 250 ng of total RNA was reverse transcribed cDNA (Target Amp labelling kit (Illumina, SanDiego, CA, USA and Superscript III, Life Technologies, Gaithersburg, MD, USA), which was further synthesized to cRNA by in vitro transcription (Superscript III, Life Technologies, Gaithersburg, MD, USA and Target Amp labelling kit, Illumina, SanDiego, CA, USA). Subsequently, unincorporated nucleotides were removed using the RNeasy kit (QIAGEN, Hilden, Germany) and the cRNA was hybridized to the Illumina Human HT-12 v4 Bead Chip according to the manufacturers' instructions using an Illumina High Scan SQ.

Pre-processing of RNA microarray data relied on the intensities of 47,323 transcripts derived from Illumina BeadStudio in 1,029 individuals of the Sorbs cohort. Steps for pre-processing comprised 1. filtering of individuals with atypical low number of expressed genes (median - 3 interquartile ranges (IQR) of the cohort's values), 2. filtering individuals with atypical log-transformed and quantile-normalized gene-expression profiles (Euclidian distance to average expression larger than median +3 IQR), and 3. filtering individuals with atypical values of internal quality parameters (quantified as Mahalanobis distance of log-transformed and quantile-normalized gene-expression data from QC probes included on the HT-12 v4 chip by Illumina, individuals having a larger value than median +3 IQR of this measure were excluded). A total of 924 individuals fulfilled all quality criteria. For 898 of these, SNP array data were also available.

Genotyping

The cohort was recruited from the self-contained Sorb population in Germany (Gross et al., 2011). The study was approved by the ethics committee of the University of Leipzig and all subjects gave written informed consent before taking part in the study.

Subjects were genotyped by either by 500 K Affymetrix GeneChip or Affymetrix Genome-Wide Human SNP Array 6.0. The BRLMM algorithm (Affymetrix, Inc) was applied for the 500 K array and the Birdseed algorithm was applied for the Genome-Wide Human SNP Array 6.0. QC of samples was performed as described in Gross et al., 2011, resulting in N=977 individuals with genotypes of good quality (N=483 genotyped with the 500 K assay, N=494 genotyped with the 6.0 assay). Genotype imputation was performed separately for individuals genotyped with the two different assays. No prior SNP filtering was performed. Imputation was done with IMPUTE v2.1.2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) using HapMap2 CEU, Release 24, dbSNP-build 126, NCBI 36 as the reference panel. For post-imputation QC, SNPs with minor allele frequency<1% or info-score<0.5 were removed. To detect ethnical outliers, a ‘drop one in’ procedure was done to avoid bias by the relatedness structure within the Sorbs cohort (Veeremah et al., 2011). PCA was performed for each Sorbian individual together with the 50 most unrelated HapMap CEU individuals based on genotype data as explained in Gross et al., 2011. Resulting eigenvectors of CEU individuals were averaged over all iterations. Individuals were considered as ethnical outliers if the distance from the mean of the respective eigenvector of at least one of the first 10 eigenvectors exceeds 6 sd. After application of filtering, three individuals were discarded from association analysis, leaving 824 for further analysis.

For eQTL analysis, individuals with both genotyping and expression data were filtered. Additionally, individuals with relatedness greater than 0.2 were removed (Wang et al., 2002). After validating that no mix-up of gene expression and genetic data is present, a total of 513 individuals were included in *cis*-eQTL and *trans*-eQTL analysis.

References

Gross, A., Tönjes, A., Kovacs, P., Veeramah, K. R., Ahnert, P., Roshyara, N. R., … Scholz, M. (2011). Population-genetic comparison of the Sorbian isolate population in Germany with the German KORA population using genome-wide SNP arrays. *BMC Genetics*, *12*(1), 67. http://doi.org/10.1186/1471-2156-12-67

Tönjes, A., Koriath, M., Schleinitz, D., Dietrich, K., Böttcher, Y., Rayner, N. W., … Stumvoll, M. (2009). Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs. *Human Molecular Genetics*, *18*(23), 4662–4668. http://doi.org/10.1093/hmg/ddp423

Veeramah, K. R., Tönjes, A., Kovacs, P., Gross, A., Wegmann, D., Geary, P., … Stumvoll, M. (2011). Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *European Journal of Human Genetics*, *19*(9), 995–1001. http://doi.org/10.1038/ejhg.2011.65

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, *160*(3), 1203–15.

Acknowledgments

We thank Knut Krohn and Beate Enigk for conducting microarray experiments of the Sorbs sample at the IZKF Leipzig at the Faculty of Medicine of the University of Leipzig (Projekt Z03). This work was supported by grants from the German Research Foundation (SFB-1052 ‘‘Obesity mechanisms’’ A01, B03, SPP 1629 TO 718/2- 1), from the German Diabetes Association, from the DHFD (Diabetes Hilfs- und Forschungsfonds Deutschland) and from IFB Adiposity Diseases (AD2-060E, AD2-06E95, AD2-06E99). IFB Adiposity Diseases is supported by the Federal Ministry of Education and Research (BMBF), Germany, FKZ: 01EO1501.

YFS

The Cardiovascular Risk in Young Finns Study (YFS) is a population-based, prospective multi-center cohort study being conducted in five university hospital cities in Finland (Raitakari et al., 2008). Of the 3,596 individuals participating at baseline in 1980, 2,050 individuals took part in the YFS follow-up examinations in 2011-2012. As described previously (Elovainio et al., 2015), 2,049 of them gave blood samples for RNA isolation. Of these, 1,664 samples had high enough concentration after the amplification step and were analyzed with Illumina HumanHT-12 version 4 Expression BeadChips.

In YFS, 2.5 mL of whole blood was collected into PaXgene Blood RNA Tubes (PreAnalytix, Hombrechtikon, Switzerland). Each tube was inverted 8-10 times and then stored at room temperature for at least 2 h. PaXgene tubes were frozen and stored for <1 year at −80°C. After thawing, tubes were stored at room temperature for 2–12h, following the manufacturer’s instructions. RNA was then isolated with the PAXgene Blood RNA Kit (Qiagen) with the DNase Set according to manufacturer’s instructions. We used a QiaCube isolation robot. The concentrations and purity of the RNA samples were evaluated spectrophotometrically with NanoDrop (BioPhotomer, Eppendorf, Wesseling-Berzdorf, Germany). We reverse-transcribed 200 ng of RNA into cDNA and biotin-UTP-labeled using the Illumina TotalPrep RNA Amplification Kit (Ambion); 1500 ng of cDNA was then hybridized to the Illumina HumanHT-12 v4 Expression

BeadChip (Illumina Inc.). The BeadChips were scanned with the Illumina HiScan system. We exported raw Illumina probe data from Beadstudio, combined the data files using the limma R package, and extracted the raw expression data for the eQTLGen consortium analysis pipeline. After sample mix-up correction, both genotype and gene expression data were available for 1,428 individuals. The genotyping was done using a custom-built Illumina Human 670 k BeadChip at the Wellcome Trust Sanger Institute. Before imputation, QC was performed for the genotype data and SNPs with Hardy–Weinberg *p*-value of $<1\times10^{-5}$ were excluded. Genotype imputation was performed using IMPUTE2 (Howie et al., 2012) and 1000 Genomes Phase I Integrated Release Version 3 (Mar 2012) samples as a reference. (Raitoharju et al., 2014; Turpeinen et al., 2015).

References

Elovainio, M., Taipale, T., Seppälä, I., Mononen, N., Raitoharju, E., Jokela, M., … Lehtimäki, T. (2015). Activated immune–inflammatory pathways are associated with long-standing depressive symptoms: Evidence from gene-set enrichment analyses in the Young Finns Study. *Journal of Psychiatric Research*, *71*, 120–125. http://doi.org/10.1016/j.jpsychires.2015.09.017

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. http://doi.org/10.1038/ng.2354

Raitakari, O. T., Juonala, M., Ronnemaa, T., Keltikangas-Jarvinen, L., Rasanen, L., Pietikainen, M., … Viikari, J. S. (2008). Cohort Profile: The Cardiovascular Risk in Young Finns Study. *International Journal of Epidemiology*, *37*(6), 1220–1226. http://doi.org/10.1093/ije/dym225

Raitoharju, E., Seppälä, I., Oksala, N., Lyytikäinen, L.-P., Raitakari, O., Viikari, J., … Lehtimäki, T. (2014). Blood microRNA profile associates with the levels of serum lipids and metabolites associated with glucose metabolism and insulin resistance and pinpoints pathways underlying metabolic syndrome. *Molecular and Cellular Endocrinology*, *391*(1–2), 41–49. http://doi.org/10.1016/j.mce.2014.04.013

Turpeinen, H., Seppälä, I., Lyytikäinen, L.-P., Raitoharju, E., Hutri-Kähönen, N., Levula, M., … Pesu, M. (2015). A genome-wide expression quantitative trait loci analysis of proprotein convertase subtilisin/kexin enzymes identifies a novel regulatory gene variant for FURIN expression and blood pressure. *Human Genetics*, *134*(6), 627–636. http://doi.org/10.1007/s00439-015-1546-5

Acknowledgments

The Young Finns Study has been financially supported by the Academy of Finland: grants 286284, 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility area of Kuopio, Tampere and Turku University Hospitals (grant X51001); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research ; Finnish Cultural Foundation; The Sigrid Juselius Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahnsson Foundation; Signe and Ane Gyllenberg Foundation; Diabetes Research Foundation of Finnish Diabetes Association; and EU Horizon 2020 (grant 755320 for TAXINOMISIS); and European Research Council (grant 742927 for MULTIEPIGEN project); Tampere University Hospital Supporting Foundation. We thank the teams that collected data at all measurement time points; the persons who participated as both children and adults in these longitudinal studies; and biostatisticians Irina Lisinen, Johanna Ikonen, Noora Kartiosuo, Ville Aalto, and Jarno Kankaanranta for data management and statistical advice.

RNA-seq cohorts

BIOS Consortium

The Biobank-based Integrative Omics Study (BIOS, http://www.bbmri.nl/acquisition-use-analyze/bios/) Consortium has been set up in an effort of several Dutch biobanks to create a homogenized dataset with different levels of 'omics' data layers. Genotyping was performed in each cohort separately, as described before: LifeLines DEEP (LLD; Tigchelaar et al., 2015), Leiden Longevity Study (LLS; Schoenmaker et al., 2005; Deelen et al., 2016), Netherlands Twin Registry (NTR; Lin et al., 2016); Rotterdam Study (RS; Hofman et al., 2013; Hofman et al., 2015), Prospective ALS Study Netherlands (PAN; Huisman et al., 2011). All genotypes were imputed to the Haplotype Reference Consortium (HRC, McCarthy et al., 2016) using the Michigan imputation server (Das et al., 2016).

RNA-seq gene expression data was generated in The Human Genotyping facility (HugeF, Erasmus MC, Rotterdam, the Netherlands, http://www.blimdna.org). RNA-seq extraction and processing has been described before for a subset of the data (Zhernakova et al., 2017). Briefly, RNA was extracted from whole blood and paired-end sequenced using Illumina HiSeq 2000. Reads were aligned using STAR 2.3.0e (Dobin et al., 2013) while masking common (MAF > 0.01) SNPs from the Genome of the Netherlands (Genome of the Netherlands Consortium, 2014). Gene-level expression was quantified using HTseq (Anders et al., 2015). FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check quality metrics, and we removed individuals with < 70% of reads mapping to exons (exon mapped / genome). We included only unrelated individuals in this analysis removed population outliers by filtering out samples with >3 standard deviations from the average heterogeneity score. We removed 25 PCs, from the expression matrix with all cohorts combined, to account for unmeasured variation. Here, we briefly describe each cohort.

CODAM

The Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) is group of individuals with a slightly increased risk of cardiometabolic disease selected from a population-based cohort (Van Greevenbroek et al., 2011). Individuals in CODAM are of European descent and older than 40 years of age. They have either an increased BMI (>25), a family history of type 2 diabetes, previous gestational diabetes and/or glycosuria, or they use medication to treat hypertension.

LLD

LifeLines is a population-based longitudinal cohort study that includes questionnaire-based and clinical data of 167,729 individuals living in the three Northernmost provinces of the Netherlands. The study specifically focuses on families and employs a three-generational design. LLD is a subset of 1,500 unrelated LifeLines participants who consented to further investigation of their genetics, gene expression, methylation, gut microbiome and exhaled breath metabolomics.

LLS

The LLS cohort studies families with individuals that reach a high age without health problems. At least two long-lived siblings (men > 88 years, women > 90 years) were required to be alive at the time of ascertainment, and their children and grandchildren are also included in the study. A total of 944 siblings from 421 European-descent families were recruited with 1,671 of their offspring and 744 partners.

NTR

The Netherlands Twin Register was set up in 1987 (https://tweelingenregister.vu.nl) to recruit Dutch mono- and dizygotic twins and their families. The NTR investigates health and lifestyle (Willemsen et al., 2013). Twins and their relatives complete questionnaires and provide clinical measurements. From 2004 onwards, a subset of participants were asked to donate blood in order to create a biobank. Blood samples were used for genotyping, DNA and RNA isolation and to biomarker studies (Willemsen et al., 2010; Wright et al. 2014). A subset of twins is also part of the BIOS consortium, and we selected one individual from each twin pair for our study.

Acknowledgments

We are extremely grateful to the twin families who take part in NTR and to the study team. We acknowledge support from BBMRI–NL (Biobanking and Biomolecular Resources Research Infrastructure 184.021.007 and 184.033.111); Spinozapremie (NWO- 56-464-14192), the European Research Council (ERC Advanced 230374) and KNAW Academy Professor Award (PAH/6635) to DIB, the National Institutes of Health (NIH, Grand Opportunity grants 1RC2 MH089951, and 1RC2 MH089995); the Avera Institute for Human Genetics, Sioux Falls, South Dakota (USA) and NWO-Groot 480-15-001/674: Netherlands Twin Registry Repository.

RS

The Rotterdam Study has been described above. A subset of the Rotterdam Study is also part of the BIOS consortium, and these samples have been RNA-sequenced. We excluded samples that were also measured on the Illumina expression arrays.

PAN

PAN is a prospective study for patients suffering from amyotrophic lateral sclerosis (ALS). Since 2006, PAN aims to include all Dutch patients with ALS and similar phenotypes to correlate potential lifestyle, genetic and environmental risk factors with the onset and prognosis of ALS (https://www.als-centrum.nl/kennisplatform/prospectieve-als-studie-nederland-pan/). To date, 3,400 patients have been included, and genotypes and expression data have been generated for a subset of these patients.

References

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. http://doi.org/10.1093/bioinformatics/btu638

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. http://doi.org/10.1038/ng.3656

Deelen, J., van den Akker, E. B., Trompet, S., van Heemst, D., Mooijaart, S. P., Slagboom, P. E., & Beekman, M. (2016). Employing biomarkers of healthy ageing for leveraging genetic studies into human longevity. *Experimental Gerontology*, *82*, 166–174. http://doi.org/10.1016/j.exger.2016.06.013

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Hofman, A., Brusselle, G. G. O., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., … Vernooij, M. W. (2015). The Rotterdam Study: 2016 objectives and design

update. *European Journal of Epidemiology*, *30*(8), 661–708. http://doi.org/10.1007/s10654-015-0082-x

Hofman, A., Murad, S. D., van Duijn, C. M., Franco, O. H., Goedegebure, A., Arfan Ikram, M., … Vernooij, M. W. (2013). The Rotterdam Study: 2014 objectives and design update. *European Journal of Epidemiology*, *28*(11), 889–926. http://doi.org/10.1007/s10654-013-9866-z

Huisman, M. H. B., de Jong, S. W., van Doormaal, P. T. C., Weinreich, S. S., Schelhaas, H. J., van der Kooi, A. J., … van den Berg, L. H. (2011). Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *Journal of Neurology, Neurosurgery & Psychiatry*, *82*(10), 1165–1170. http://doi.org/10.1136/jnnp.2011.244939

Lin, B. D., Willemsen, G., Abdellaoui, A., Bartels, M., Ehli, E. A., Davies, G. E., … Hottenga, J. J. (2016). The Genetic Overlap Between Hair and Eye Color. *Twin Research and Human Genetics*, *19*(06), 595–599. http://doi.org/10.1017/thg.2016.85

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Consortium, for the H. R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–83. http://doi.org/10.1038/ng.3643

Schoenmaker, M., de Craen, A. J. M., de Meijer, P. H. E. M., Beekman, M., Blauw, G. J., Slagboom, P. E., & Westendorp, R. G. J. (2006). Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European Journal of Human Genetics*, *14*(1), 79–84. http://doi.org/10.1038/sj.ejhg.5201508

Tigchelaar, E. F., Zhernakova, A., Dekens, J. A. M., Hermes, G., Baranska, A., Mujagic, Z., … Feskens, E. J. M. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*, *5*(8), e006772. http://doi.org/10.1136/bmjopen-2014-006772

van Greevenbroek, M. M. J., Jacobs, M., van der Kallen, C. J. H., Vermeulen, V. M. M.-J., Jansen, E. H. J. M., Schalkwijk, C. G., … Stehouwer, C. D. A. (2011). The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *European Journal of Clinical Investigation*, *41*(4), 372–379. http://doi.org/10.1111/j.1365-2362.2010.02418.x

Willemsen, G., de Geus, E. J. C., Bartels, M., van Beijsterveldt, C. E. M. T., Brooks, A. I., Estourgie-van Burk, G. F., … Boomsma, D. I. (2010). The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. *Twin Research and Human Genetics*, *13*(03), 231–245. http://doi.org/10.1375/twin.13.3.231

Willemsen, G., Vink, J. M., Abdellaoui, A., den Braber, A., van Beek, J. H. D. A., Draisma, H. H. M., … Boomsma, D. I. (2013). The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, *16*(1), 271–81. http://doi.org/10.1017/thg.2012.140

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou YH, Abdellaoui A, Batista S, Butler C, Chen G, Chen TH, D'Ambrosio D, Gallins P, Ha MJ, Hottenga JJ, Huang S, Kattenberg M, Kochar J, Middeldorp CM, Qu A, Shabalin A, Tischfield J, Todd L, Tzeng JY, van Grootheest G, Vink JM, Wang Q, Wang W, Wang W, Willemsen G, Smit JH, de Geus EJ, Yin Z, Penninx BW, Boomsma DI.

Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014 May;46(5):430-7. doi: 10.1038/ng.2951

Zhernakova, D. V, Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., … Franke, L. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*, *49*(1), 139–145. http://doi.org/10.1038/ng.3737

EGCUT RNA-seq

EGCUT RNA-seq dataset has been described elsewhere (Lepik et al., 2017). EGCUT RNA-seq analyses were approved by Ethics Review Committee of Human Research of the University of Tartu, Estonia (permission no 234/T-12).

RNA was extracted from thawed Tempus tubes using TRIzol Reagent (Invitrogen) and further purified using RNeasy Mini Kit (Qiagen). Globin mRNA was depleted using GLOBINclear Kit (Invitrogen). RNA quality was checked using an Agilent 2200 TapeStation (Agilent Technologies). Sequencing libraries were prepared using 200 ng of RNA according to the Illumina TruSeq stranded mRNA protocol. RNA sequencing was performed at the Estonian Genome Center Core Facility using Illumina paired-end 50 bp sequencing technology according to manufacturer specifications.

We used fastQC v.0.11.3 for raw data quality control and Trimmomatic (version 0.36, Bolger et al., 2014) to remove 3 leading and 3 trailing bases and remove adapters. For adapter removal we used adapter file provided with fastQC. Additional read quality filtering was made using FASTX Toolkit v.0.0.13 fastq_quality_filter script with minimum quality score 30 and minimum 50% of base pairs with required quality.

Quality control was done by FastQC (version 0.11.2, Andrews et al., 2010). The quality filtered data was mapped on genome hg19, position sorted and indexed with STAR v. 2.5.2 (STAR index files were provided by eQTLGen, Dobin et al., 2013). The mapped data quality statistics was collected with Picardtools v.1.130 CollectRnaSeqMetrics. Read counts were obtained using HTSeq-count script v. 0.6.1 and GRCh37.v71 annotation file.

After data preprocessing and QC, 508 samples were enrolled into eQTLGen meta-analyses.

References

Lepik, K., Annilo, T., Kukuškina, V., Kisand, K., Kutalik, Z., Peterson, P., & Peterson, H. (2017). C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Computational Biology*, *13*(9), e1005766. https://doi.org/10.1371/journal.pcbi.1005766

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. http://doi.org/10.1093/bioinformatics/btu170

Andrews S. FastQC: A quality control tool for high throughput sequence data 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNAseq aligner. Bioinformatics. 2013; 29(1):15–21. Epub 2012/10/25. https://doi.org/10.1093/ bioinformatics/bts635 PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.

Acknowledgements

EGCUT analyses were funded by EU H2020 grant 692145, Estonian Research Council Grant IUT20-60, IUT24-6, and European Union through the European Regional Development Fund Project No. 2014-2020.4.01.15-0012GENTRANSMED. This work was carried out in the High Performance Computing Center of University of Tartu.

CartaGene

CARTaGENE (CaG) is a population-based cohort comprising over 40,000 women and men participants randomly recruited among the three urban centers (Montreal, Quebec city and Saguenay) of the province of Quebec, Canada. CaG is a regional cohort within the Canadian Partnership for Tomorrow Project (CPTP), which included >315,000 participants. CaG targeted the segment of the population that is most at risk of developing chronic diseases, with participants' ages ranging from 40 to 60 years old. Health and social-demographic information, such as disease history, physiological measures, lifestyle and environmental factors, were collected for each individual along with biological samples (Awadalla et al. 2013).

We selected 708 samples (set 1) from the CaG biobank based on the availability of Tempus blood RNA tubes and on the Framingham risk scores to ensure an equal distribution of ages and gender. In a second phase, 292 samples (set 2) were included based on their RNA and arterial stiffness measures availability. These samples, provided by participants with high and average values of arterial stiffness, were chosen to achieve a uniform range of arterial stiffness values.

Gene expression

Whole blood samples from participants included in set 1 and 2 were collected in 2010. Total RNA was isolated using Tempus Spin RNA isolation kit (ThermoFisher Scientific), and the GLOBINclear-Human kit (ThermoFisher Scientific) was used to perform globin mRNA-depletion. All samples displayed high quality and minimal degradation of the RNA based on a RNA Integrity Number (RIN) > 7.5. Participants' transcriptomes were obtained by RNA sequencing, for which we used paired-end libraries constructed with TruSeq RNA Sample Prep kit v2 (Illumina) with 500ng of globin-depleted total RNA. Paired-end RNA-seq libraries were inspected before sequencing according to Illumina protocols and the sequencing was performed on a HiSeq 2000 platform at the Genome Quebec Innovation Center (Montreal, Canada). Set 1 (708 samples) and set 2 (292 samples) were sequenced by using three and six samples per lane, respectively.

Genotyping

High density SNP genotyping data for 928 samples with RNA-Seq profiles passing QC thresholds were obtained by using the Illumina Omni2.5 array. Variant imputation was conducted on 968 individuals. We pre-phased the genotypes with SHAPEIT (v2.r64410) (Delaneau, Zagury, and Marchini 2013) using the default parameters, on both the autosomes and the chromosome X. We filtered variants for MAF > 1% and Hardy-Weinberg p-value > 0.0001 and used the haplotypes within IMPUTE2 (v2.2.2) (Howie et al. 2012) to perform the imputation using the 1000 Genomes Phase I integrated haplotypes (Dec 2013). We used the parameters Ne = 11418 and call thresh = 0.9. We removed variants with a call rate less than 90%, MAF > 1% and Hardy-Weinberg p-value > 0.0001. A total of 9,157,622 variants passed the filters. Of these, 8,877,297 variants were found on the autosomes and included 779,579 insertion-deletion polymorphisms (indels) (8.78%) and 8,097,718 SNPs (91.22%). 280,325 variants were found on the chromosome X, which included 28,504 indels (10,16%) and 251,821 SNPs (89.84%).

After sample pre-processing and QC, 634 samples from set 1 and 191 samples from set 2 were included to eQTLGen meta-analyses.

References

Awadalla, P., Boileau, C., Payette, Y., Idaghdour, Y., Goulet, J.-P., Knoppers, B., … CARTaGENE Project. (2013). Cohort profile of the CARTaGENE study: Quebec's

population-based biobank for public health and personalized genomics. *International Journal of Epidemiology*, *42*(5), 1285–1299. http://doi.org/10.1093/ije/dys160

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6. http://doi.org/10.1038/nmeth.2307

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. http://doi.org/10.1038/ng.2354

DGN

The Depression Genes and Networks (DGN) study includes genotyping and gene expression data from 922 European individuals (463 cases of Major Depressive Disorder and 459 controls), aged 21 to 60 years, selected from a survey research panel (Battle et al., 2014). DNA was isolated from whole blood and genotyped on the Illumina HumanOmni1-Quad BeadChip. RNA was extracted from whole blood and hemoglobin RNA was removed from each sample using GLOBINclear^{TM}Kit (Invitrogen). RNA sequencing was performed using Illumina HiSeq 2000 (50bp single-ended reads) following the Illumina TruSeq RNA protocol. Reads were aligned to the NCBI v37 human reference genome using TopHat. Gene expression was quantified by HTSeq using uniquely mapped reads. Sample collection, QC, and data processing are described in detail in (Battle et al., 2014). After preprocessing and QC, 919 samples were added into eQTLGen analyses.

References

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., … Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, *24*(1), 14–24. http://doi.org/10.1101/gr.155192.113

Acknowledgements

For the DGN data, we gratefully acknowledge the resources supported by National Institutes of Health/National Institute of Mental Health grants 5RC2MH089916 (PI: Douglas F. Levinson, M.D.; Co- investigators: Myrna M. Weissman, Ph.D., James B. Potash, M.D., MPH, Daphne Koller, Ph.D., and Alexander E. Urban, Ph.D.) and 3R01MH090941 (Co-investigator: Daphne Koller, Ph.D.).

GTEx

The Genotype-Tissue Expression Consortium (GTEx) v6p release (GTEx Consortium, 2017) contained samples from 44 healthy tissues of 20-70 year old human postmortem donors. DNA isolated from whole blood samples were genotyped with Illumina HumanOmni 2.5M and 5M arrays. There were ~2.2 million variants common between the two platforms. An additional ~12.5 million variants were imputed with IMPUTE2 using the multi-ethnic reference panel from 1000 Genomes Project Phase 1 v3. RNA-seq was performed for samples with a minimum RIN score of 5.7 and a minimum of 500 ng of total RNA using either Illumina HiSeq 2000 or Illumina HiSeq 2500 (76bp paired-end reads) following the Illumina TrueSeq RNA protocol. Reads were aligned to the human reference genome GRCh37/hg19 with Tophat v1.4.1. Gene expression was quantified as reads per kilobase of transcript per million mapped reads (RPKM) by RNA-SeQC with -strictMode flag based on the GENCODE Release 19 annotation using uniquely mapped,

properly paired reads contained fully within exon boundaries and with maximum alignment distance of 6. 338 individuals had both genotyping and RNA sequencing data from whole blood. Sample collection, QC, and data processing are described in detail in (GTEx Consortium, 2017).

References

GTEx Consortium, Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., … Zhu, J. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. http://doi.org/10.1038/nature24277

Acknowledgements

The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH). Additional funds were provided by the National Cancer Institute; National Human Genome Research Institute (NHGRI); National Heart, Lung, and Blood Institute; National Institute on Drug Abuse; National Institute of Mental Health; and National Institute of Neurological Disorders and Stroke. Donors were enrolled at the Biospecimen Source Sites funded by Leidos Biomedical, Inc. (Leidos) subcontracts to the National Disease Research Interchange (10XS170) and Roswell Park Cancer Institute (10XS171). The LDACC was funded through a contract (HHSN268201000029C) to The Broad Institute. Biorepository operations were funded through a Leidos subcontract to the Van Andel Institute (10ST1035). Additional data repository and project management were provided by Leidos (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami grant DA006227.

A.Ba. was supported by NIH grant 1R01MH109905, NIH grant R01HG008150 (NHGRI; Non-Coding Variants Program), and NIH grant R01MH101814 (NIH Common Fund; GTEx Program).

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: dbGaP accession number phs000424.v7.p2 on 09/27/2018.

Affymetrix array cohorts

FHS

The Framingham Heart Study (FHS) is a cohort study initiated in 1948, with the aim of identifying risk factors for heart disease. Starting in 1971, the offspring and offspring spouses (N=5,124) of the original FHS cohort participants were recruited (Feinleib et al., 1975), and they have been examined approximately every 4 years since. From 2002 to 2005, adult children (third generation cohort, N=4,095) of the offspring cohort participants were recruited (Splansky et al., 2007) and are also being examined in an ongoing manner. For this study, we included a total of 5,075 participants from the offspring (N = 2,119) and third-generation (N = 2,956) cohorts who provided both genotype and gene expression information (Huan et al., 2015). Whole blood samples were collected at the eighth examination of the offspring cohort and the second examination of the third generation cohort. Fasting peripheral whole blood samples (2.5 ml) were stored in PAXgene^{TM} tubes (PreAnalytiX, Hombrechtikon, Switzerland) and the Affymetrix Human ExonArray ST 1.0 (Affymetrix, Inc., Santa Clara, CA) was utilized to measure mRNA expression levels. Genotyping was performed with the Affymetrix 500K mapping array and the Affymetrix 50K gene-focused MIP array. Genotype imputation was conducted using impute2 against 1000 Genomes Phase 3 reference.

***Trans*-eQTL detection pipeline for Framingham Heart Study**

Imputation results were converted to bgen format with genotype dosage as independent variables. A genetic-relatedness matrix was constructed using all of the imputed genotypes using GEMMA (Zhou et al., 2012). For the gene expression data, the first 20 non-genetic PCs of gene expression were regressed out, and residuals were used as adjusted phenotypes for the association studies.

Prior to *trans*-eQTL detection, *cis*-eQTLs were first detected by stepwise sequential conditional analysis. Variants located within a distance of less than 1 Mb from either the 5' or 3' ends of the gene coding region being explored were deemed to be *cis*, and only SNPs with a minor allele frequency (MAF) >0.01 and a HWE P-value >0.001 were included in the analyses. In each iteration of the conditional analysis, the peak signal with a P-value $<10^{-6}$, also computed using the GEMMA package, was determined to be an independent eSNP. The residuals from discovery of each SNP were then taken forward as the dependent variable in a new scan for additional independent SNP(s).

After *cis*-eQTL detection, residuals removing all *cis*-eQTL effects at each gene were used as the phenotype, and *trans*-eQTL detection was performed on 10,562 variants associated with phenotype traits. This signal was evaluated with a mixed linear model in GEMMA, controlling for population structure and relatedness. To obtain the null distribution by permutation, the covariance component was first estimated by REML, and the square root of the covariance matrix was used to transform the phenotype and genotype matrices (Abney et al., 2002). After this step, the transformed phenotype is exchangeable and can be safely used to conduct permutation analysis. The false discovery rate (FDR) was controlled relative to 10 phenotype permutations, retaining the co-expression structure by permuting the sample IDs which were used in common for all expression phenotypes.

References

Abney, M., Ober, C., & McPeek, M. S. (2002). Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *American Journal of Human Genetics*, *70*(4), 920–34. http://doi.org/10.1086/339705

Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M., & Castelli, W. P. (1975). The framingham offspring study. Design and preliminary data. *Preventive Medicine*, *4*(4), 518–525. https://doi.org/10.1016/0091-7435(75)90037-7

Huan, T., Liu, C., Joehanes, R., Zhang, X., Chen, B. H., Johnson, A. D., Yao, C., Courchesne, P., O'Donnell, C. J., Munson, P. J., & Levy, D. (2015). A systematic heritability analysis of the human whole blood transcriptome. *Human Genetics*, *134*(3), 343–358. https://doi.org/10.1007/s00439-014-1524-3

Splansky, G. L., Corey, D., Yang, Q., Atwood, L. D., Cupples, L. A., Benjamin, E. J., D'Agostino, R. B., Fox, C. S., Larson, M. G., Murabito, J. M., O'Donnell, C. J., Vasan, R. S., Wolf, P. A., & Levy, D. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, recruitment, and initial examination. *American Journal of Epidemiology*, *165*(11), 1328–1335. https://doi.org/10.1093/aje/kwm021

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. http://doi.org/10.1038/ng.2310

NTR-NESDA

Subjects for eQTL analysis

The two parent projects that supplied data for the eQTL analysis are large-scale longitudinal studies: the Netherlands Study of Depression and Anxiety (NESDA) (Penninx et al., 2008) and the Netherlands Twin Register (NTR) (Boomsma et al., 2006). NESDA and NTR were both approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam (institutional review board [IRB] number IRB-2991 under Federal wide Assurance 3703; IRB/institute codes: NESDA 03-183 and NTR 03-180). All participants provided written informed consent. The samples used for eQTLGen analyses consisted of 2,767 subjects of European ancestry (1,901 unrelated NESDA subjects and 866 unrelated NTR subjects). The age of the participants ranged from 17 to 88 years (mean=38, *SD*=13), and 65% of the sample was female. The data used for this study overlaps with that used in our earlier studies (Jansen et al., 2017; Wright et al., 2014). For NESDA, the same samples were used in both studies. For NTR, only unrelated samples that were not present in the BIOS eQTL sample set used for the eQTLGen meta-analysis were used. Population stratification was corrected in the *cis*-eQTL, *trans*-eQTL and eQTS analyses by regressing out the three first genotype PCs from the gene expression matrix.

Blood sampling, RNA extraction, and RNA expression measurement

Study protocols and biological sample collection methods were harmonized between NTR and NESDA. RNA processing and measurements have been described in detail previously (Wright et al., 2014; Jansen et al., 2014). Venous blood samples were drawn in the morning after an overnight fast. Heparinized whole blood samples were transferred within 20 minutes of sampling into PAXgene Blood RNA tubes (Qiagen, Valencia, California, USA) and stored at−20°C. Gene expression assays were conducted at the Rutgers University Cell and DNA Repository. Samples were hybridized to Affymetrix U219 arrays (Affymetrix, Santa Clara, CA) containing 530,467 probes summarized in 49,293 probe sets. Array hybridization, washing, staining and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol. Gene expression data were required to pass standard Affymetrix QC metrics (Affymetrix expression console) before further analysis. We excluded probes that did not map uniquely to the hg19 (Genome Reference Consortium Human Build 37) reference genome sequence from further analysis, as well as probes targeting a messenger RNA (mRNA) molecule resulting from transcription of a DNA sequence containing a SNP (based on the dbSNP137 common database). After this filtering step, data for analysis remained for 423,201 probes, which could be summarized into probe sets targeting 18,238 genes. Normalized probe set expression values were obtained using Robust Multi-array Average (RMA) normalization as implemented in the Affymetrix Power Tools software (APT, version 1.12.0, Affymetrix). Data for samples that displayed a low average Pearson correlation with the probe set expression values of other samples and samples with incorrect sex-chromosome expression were removed, leaving 2,767 subjects for analysis.

DNA extraction and SNP genotyping and imputation

DNA was extracted from peripheral blood as described previously (Boomsma et al., 2008). SNP genotype pre-imputation QC, haplotype phasing and 1000 Genomes phase 1 imputation were performed as described previously (Nivard et al., 2014). Imputed SNP genotypes were coded into the reference allele dosage format and filtered at MAF >0.01 and HW P-value $>1\times10^{-4}$.

RNA processing

RNA processing was done using the following normalization steps: RMA normalization, Z-transformation, removal of the first three PCs from the imputed genotype information to correct for population stratification and removal of the first 20 non-genetic PCs from the expression data.

References

Boomsma, D. I., de Geus, E. J. C., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J.-J., … Willemsen, G. (2006). Netherlands Twin Register: From Twins to Twin Families. *Twin*

Research and Human Genetics, *9*(6), 849–857. http://doi.org/10.1375/183242706779462426

Boomsma, D. I., Willemsen, G., Sullivan, P. F., Heutink, P., Meijer, P., Sondervan, D., … Penninx, B. W. J. H. (2008). Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. *European Journal of Human Genetics*, *16*(3), 335–342. http://doi.org/10.1038/sj.ejhg.5201979

Jansen, R., Batista, S., Brooks, A. I., Tischfield, J. A., Willemsen, G., van Grootheest, G., … Penninx, B. W. (2014). Sex differences in the human peripheral blood transcriptome. *BMC Genomics*, *15*(1), 33. http://doi.org/10.1186/1471-2164-15-33

Jansen, R., Hottenga, J.-J., Nivard, M. G., Abdellaoui, A., Laport, B., de Geus, E. J., … Boomsma, D. I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics*, *26*(8), 1444–1451. http://doi.org/10.1093/hmg/ddx043

Nivard, M. G., Mbarek, H., Hottenga, J. J., Smit, J. H., Jansen, R., Penninx, B. W., … Boomsma, D. I. (2014). Further confirmation of the association between anxiety and CTNND2: replication in humans. *Genes, Brain and Behavior*, *13*(2), 195–201. http://doi.org/10.1111/gbb.12095

Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., … NESDA Research Consortium. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, *17*(3), 121–140. http://doi.org/10.1002/mpr.256

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., … Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, *46*(5), 430–437. http://doi.org/10.1038/ng.2951

Replication cohorts

LCL cohorts

ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a prospective birth cohort that recruited pregnant women with expected delivery dates between April 1991 and December 1992 from Bristol, UK. 14,541 pregnant women were initially enrolled and 14,062 children born. Detailed information on health and development of the children and their parents were collected from regular clinic visits and completion of questionnaires. A detailed description of the cohort is available on our website (http://www.bristol.ac.uk/alspac/researchers/) and has been published previously (Boyd et al., 2013). The study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

Gene expression data was generated as described previously (Bryois et al., 2014). Briefly, lymphoblastoid cell lines were established by transforming lymphocytes from blood samples taken when the study participants were 9 years old, using Epstein Barr Virus. Lymphoblastoid cell lines (LCLs) from unrelated individuals were grown under identical conditions, and cells frozen in RNAlater. RNA was extracted using an RNeasy extraction kit (Qiagen) and amplified using the

Illumina TotalPrep-96 RNA Amplification kit (Ambion). Expression profiling of the samples, each with two technical replicates, were performed using the Illumina Human HT-12 V3 BeadChips (Illumina Inc) including 48,804 l probes where 200 ng of total RNA was processed according to the Illumina protocol. Raw data was imported into the Illumina Beadstudio software, and probes with less than three beads present were excluded. Log2-transformed expression signals were then normalized with quantile normalization of the replicates of each individual, followed by quantile normalization across all individuals. We restricted our analysis to 23,935 probes tagging genes annotated in Ensembl.

Data processing, QC, and *cis*-eQTL, *trans*-eQTL and eQTS analyses were conducted using the eQTLGen analysis plan for Illumina arrays. Independent *cis*-eQTL effects were removed from the expression matrix prior to *trans*-eQTL and eQTS analyses. Results of summary-statistic-based conditional *cis*-eQTL analysis from 14,115 eQTLGen blood samples (profiled by Illumina arrays) were used for this correction. After preprocessing, 867 samples were enrolled in the eQTLGen analyses.

References

Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, Molloy L, Ness A, Ring S, Davey Smith G. Cohort Profile: The 'Children of the 90s'; the index offspring of The Avon Longitudinal Study of Parents and Children (ALSPAC). International Journal of Epidemiology 2013; 42: 111-127.

Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P, Davey Smith G, Dermitzakis ET. Cis and trans effects of human genomic variants on gene expression. PLoS Genet. 2014 Jul 10;10(7):e1004461. doi: 10.1371/journal.pgen.1004461

Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. International Journal of Epidemiology 2013; 42:97- 110.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

The UK Medical Research Council and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This research was specifically funded by ERC 260927, Swiss National Foundation 130342 and Wellcome Trust and MRC 092731. GH works in a unit that receives funding from the UK MRC (MC_UU_12013/1&2&5) and the University of Bristol.

CoLaus

The Cohorte Lausannoise (CoLaus) study is a population-based cross-sectional study of 6,188 participants residing in Lausanne, Switzerland, who were first enrolled from 2003 to 2006 (Firmann et al., 2008). The main aims of the study are to obtain a deeper knowledge of the epidemiology of cardiovascular diseases and to discover new genetic determinants of cardiovascular risk factors. LCLs were derived from blood samples. For 5,435 subjects, nuclear DNA was extracted from the LCLs for SNP genotyping using the Affymetrix GeneChip Human

Mapping 500K array set and the BRLMM genotype calling method (Affymetrix, 2006). 375k SNPs were successfully genotyped and passed the pre-imputation QC consisting of minor allele frequency >1%, call rate >90% and Hardy Weinberg Equilibrium P-value >1E-6. The pre-phasing was done using Shapeit2 (Delaneau et al., 2014), followed by imputation by minimac3 against the Haplotype Reference Consortium panel (HRC r1.1 (McCarthy et al., 2016)) hosted on the Michigan Imputation Server (Das et al., 2016). For 555 subjects, gene expression profiles of LCLs were obtained using the Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA).

cis-eQTL, *trans*-eQTL and eQTS analyses were conducted using eQTLGen analysis plan for RNA-seq datasets.

References

BRLMM : an Improved Genotype Calling Method for the GeneChip ® Human Mapping 500 K Array Set - Semantic Scholar. (n.d.). Retrieved July 26, 2018, from https://www.semanticscholar.org/paper/BRLMM-%3A-an-Improved-Genotype-Calling-Method-for-the/b113d9ec5ce87a4597aa868ccc043e7881f2d224

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. http://doi.org/10.1038/ng.3656

Delaneau, O., Marchini, J., Consortium, T. 1000 G. P., McVean, G. A., Donnelly, P., Lunter, G., … Peltonenz, L. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, *5*(1), 3934. http://doi.org/10.1038/ncomms4934

Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pécoud, A., Hayoz, D., … Vollenweider, P. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders*, *8*(1), 6. http://doi.org/10.1186/1471-2261-8-6

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Consortium, for the H. R. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–83. http://doi.org/10.1038/ng.3643

Acknowledgements

S.B. was supported by the Swiss National Science Foundation (310030-152724). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

MuTHER

The Multiple Tissue Human Expression Resource (MuTHER) LCL dataset has been described in detail before (Buil et al., 2015). 762 individuals were enrolled in this replication analysis, consisting of 134 monozygotic twins, 192 dizygotic twins and 113 unrelated individuals. The effective sample size of 535 (N=134+1.5×192+113) was used as a weight in the subsequent weighted Z-score meta-analysis.

MuTHER RNA-seq data were mapped to the GRCh37 reference genome (Lander et al., 2001) using GEM version 1.7.1 (Marco-Sola et al., 2012), and genes were quantified to RPKM values using the GENCODE 19 annotation (Harrow et al., 2012). RPKM values were scaled and centered and data were then mapped to a normal distribution.

For replication analyses, all eQTLGen hits were compared to the matched gene-variant pair in MuTHER dataset. The model was fitted together with 14 PCs (not showing any Bonferroni significant GWAS association, 0.05/6263243 SNPs, calculated with GEMMA), *cis*-eQTLs identified from the same dataset (FDR < 0.05) and family structure, using lmer R package and random effects.

References

Buil, A., Brown, A. A., Lappalainen, T., Viñuela, A., Davies, M. N., Zheng, H.-F., … Dermitzakis, E. T. (2015). Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nature Genetics*, *47*(1), 88–91. http://doi.org/10.1038/ng.3162

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–74. http://doi.org/10.1101/gr.135350.111

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, *9*(12), 1185–1188. http://doi.org/10.1038/nmeth.2221

Geuvadis

The production of genotypes and RNA-seq data from the publicly available Geuvadis dataset has been described previously (Lappalainen et al., 2013). We downloaded the data, harmonized the genotypes to 1000G p1v3 using Genotype Harmonizer, and processed the RNA-seq information using the steps outlined in the **Methods**. In brief, we TMM-normalized, CPM-filtered, log2-normalized, Z-transformed and removed the first 20 non-genetic PCs. We included the individuals from four European populations: CEPH (CEU), Finns (FIN), British (GBR), and Toscani (TSI), and meta-analyzed them as separate cohorts.

References

Lappalainen, T., Sammeth, M., Friedländer, M. R., ‘t Hoen, P. A. C., Monlong, J., Rivas, M. A., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. http://doi.org/10.1038/nature12531

LCL replication meta-analysis

The results from LCL replication cohorts were meta-analyzed using weighted Z-score method, using the same approach used in the discovery meta-analysis (**Methods**) and including all available summary statistics. As summary statistics from permuted analyses were not available for all the datasets, we applied Benjamini-Hochberg method to correct for multiple testing.

iPSC datasets

i2QTL Consortium

The i2QTL Consortium was set up to jointly analyze five induced pluripotent stem cell (iPSC) datasets (HipSci (Kilpinen et al., 2017), iPSCORE (Panopoulous et al., 2017), GENESiPS (Carcamo-Orive et al., 2017), PhiLiPS (Pashos et al., 2017) and the Banovich study (Banovich et al., 2018). This integrative analysis allowed to identify iPSC-specific genetic regulation of gene expression in both a *cis* and *trans* setting. Genotyping and RNA-sequencing-based gene expression profiling was performed in each cohort separately, as described in detail before.

In the i2QTL Consortium, both genetics and RNA-seq-based gene expression data was jointly reprocessed. Raw chip genotypes from the studies were imputed using a combined imputation panel based on UK10K and 1000G (phase 1) using sample-specific imputation as described in Kilpinen et al. Raw RNA-seq reads were trimmed from their adapters and trimmed from low quality bases using Trim Galore! (Krueger et al, Martin et al, Simon et al). Trimmed reads were mapped to the human reference genome build 37 using STAR (Dobin et al., 2013) (version: 020201) in two-pass alignment mode, using the defaults proposed by the ENCODE consortium (STAR manual, Dobin et al., 2013). Based on the STAR alignment, mRNA abundance was quantified using featureCounts (Liao et al., 2014) (v1.6.0). FeatureCounts was run on the primary alignments only using the "-B" and "-C" options in stranded mode, Genome reference and transcript information was retrieved from ENSEMBL 75 (Zerbino et al., 2018). The featureCounts quantifications per sample were merged and subsequently normalized to generate edgeR (Robinson et al., 2010) normalized transcripts per kilobase million (TPM) values.

Low quality RNA-seq samples were removed leaving 1,178 iPSC lines derived from 762 European donors for analysis, all of which also have genetic information available. More information on genotyping, expression quantification and sample QC can be found in Bonder et al (Bonder et al., 2019).

To correct for multiple lines per donor and the family structure in the data, we used a linear mixed model set up as available in LIMIX (Lippert et al., 2014). Based on the LIMIX framework, a QTL pipeline matched closely to the main analysis pipeline was developed to map *cis*-, *trans*-QTL and eQTS. We corrected for population structure and 50 PEER factors were used as covariates. There was no minor allele or Hardy-Weinberg filtering and other settings, were matched to the main analysis as much as possible. The gene expression was forced into a normal distribution. To correct for multiple testing, we applied Benjamini-Hochberg FDR over the number of features tested. *Cis*-eQTLs were regressed out from the gene expression matrix before replicating *trans*-eQTLs and eQTS.

References

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., … Gilad, Y. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Research*, *28*(1), 122–131. https://doi.org/10.1101/gr.224436.117

Bonder, M. J., Smail, C., Gloudemans, M. J., Frésard, L., Jakubosky, D., D'Antonio, M., … Stegle, O. (2019). Systematic assessment of regulatory effects of human disease variants in pluripotent cells. *BioRxiv*, 784967. https://doi.org/10.1101/784967

Carcamo-Orive, I., Hoffman, G. E., Cundiff, P., Beckmann, N. D., D'Souza, S. L., Knowles, J. W., … Lemischka, I. (2017). Analysis of Transcriptional Variability in a Large Human

iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. *Cell Stem Cell*, *20*(4), 518-532.e9. https://doi.org/10.1016/j.stem.2016.11.005

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., … Gaffney, D. J. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*, *546*(7658), 370–375. http://doi.org/10.1038/nature22403

Krueger, F. (2012). Trim Galore! Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. http://doi.org/10.1093/bioinformatics/btt656

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. *BioRxiv*, 003905. http://doi.org/10.1101/003905

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. http://doi.org/10.14806/ej.17.1.200

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, *32*(10), 1479–1485. http://doi.org/10.1093/bioinformatics/btv722

Panopoulos, A. D., D'Antonio, M., Benaglio, P., Williams, R., Hashem, S. I., Schuldt, B. M., … Frazer, K. A. (2017). iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Reports*, *8*(4), 1086–1100. http://doi.org/10.1016/j.stemcr.2017.03.012

Pashos, E. E., Park, Y. S., Wang, X., Raghavan, A., Yang, W., Abbey, D., … Musunuru, K. (2017). Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell Stem Cell*, *20*(4), 558-570.e10. https://doi.org/10.1016/j.stem.2017.03.017

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–40. http://doi.org/10.1093/bioinformatics/btp616

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. http://doi.org/10.1093/nar/gkx1098

Acknowledgements

We thank i2QTL Consortium and the individual cohorts (HipSci, iPSCORE, GENESiPS, PhiLiPS) for providing the iPSC replication results.

Blood cell types

Datasets from purified blood cell types

Unpublished eQTL data from CD4+, CD8+, CD14+, CD15+ and CD19+ immune cells, macrophages and platelets from multiple sources were used to replicate the *trans*-eQTL results.

CD14+ and CD19+ data was used from (Fairfax et al. 2012) and (Fairfax et al., 2014). CD4+, CD8+, CD14+, CD15+ and CD19+, and platelet data was used from (Momozawa et al., 2018). Additionally, CD16+ neutrophil data (Naranbhai et al., 2015) was combined and analysed together with abovementioned CD15+ neutrophil data from Momozawa et al., 2018. Additional samples from CD14+ and macrophage data was used from the Cardiogenics consortium (Rotival et al., 2011). An additional unpublished platelet dataset was generated for the samples from NIHR Cambridge BioResource (http://www.cambridgebioresource.org.uk). Donors for those datasets were recruited with informed consent (REC 12/EE/0040) at the NHS Blood and Transplant, Cambridge. Overview of the samples included from each dataset is shown below.

Cell type	Original publication/Source	# samples used
CD14+	Momozawa et al., 2018	301
CD14+	Rotival et al., 2011	758
CD14+	Fairfax et al. 2012 and Fairfax et al. 2014	421
CD15+	Momozawa et al., 2018	303
CD16+	Naranbhai et al., 2015	109
CD19+	Momozawa et al., 2018	298
CD19+	Fairfax et al. 2012	285
CD4+	Momozawa et al., 2018	309
CD8+	Momozawa et al., 2018	304
Platelets	Momozawa et al., 2018	236
Platelets	NIHR Cambridge BioResource	152
Macrophages	Rotival et al., 2011	599

In all datasets, gene expression was profiled by Illumina HT12v4 and Illumina Human-Ref-8 v3. All datasets were individually reprocessed in a unified pipeline based on lumi (Du et al. 2008) and COMBAT (Leek et al. 2012) to generate a homogeneous dataset. Ten PEER (Stegle et al. 2012) factors were calculated for every dataset. Expression and genotype data was merged within cell types to increase sample size.

Prior to *trans*-eQTL replication analysis, all independent *cis*-eQTL effects were regressed out from the expression matrices. A linear mixed model from LIMIX v0.8.5 (https://github.com/limix/limix) was used to run *trans*-eQTL analyses by testing all 10,562 trait-associated SNPs against all the available genes. Population structure (normalized genotype covariance) was taken into account and 10 PEER factors were used as covariates in *cis*- and *trans*-eQTL mapping.

References

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O., & Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics*, *44*(5), 502–510. https://doi.org/10.1038/ng.2205

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., & Knight, J. C. (2014). Innate immune activity conditions the

effect of regulatory variants upon monocyte gene expression. *Science (New York, N.Y.)*, *343*(6175), 1246949. https://doi.org/10.1126/science.1246949

Rotival, M., Zeller, T., Wild, P. S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., Godefroy, T., Perret, C., Germain, M., Eleftheriadis, M., Sinning, C. R., Schnabel, R. B., Lubos, E., Lackner, K. J., Rossmann, H., … Blankenberg, S. (2011). Integrating Genome-Wide genetic variations and monocyte expression data reveals Trans-Regulated gene modules in humans. *PLoS Genetics*, *7*(12), 1002367. https://doi.org/10.1371/journal.pgen.1002367

Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., Crins, F., Docampo, E., Elansary, M., Gori, A. S., Lecut, C., Mariman, R., Mni, M., Oury, C., Altukhov, I., Alexeev, D., Aulchenko, Y., Amininejad, L., Bouma, G., … Zhao, Z. Z. (2018). IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nature Communications*, *9*(1), 1–18. https://doi.org/10.1038/s41467-018-04365-8

Naranbhai, V., Fairfax, B. P., Makino, S., Humburg, P., Wong, D., Ng, E., Hill, A. V. S., & Knight, J. C. (2015). Genomic modulators of gene expression in human neutrophils. *Nature Communications*, *6*(1), 1–13. https://doi.org/10.1038/ncomms8545

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547–1548. https://doi.org/10.1093/bioinformatics/btn224

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883. https://doi.org/10.1093/bioinformatics/bts034

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Acknowledgements

Research in the Ouwehand laboratory receives funding from the British Heart Foundation, European Commission (TrainMALTA), International Society on Thrombosis and Haemostasis, National Institute for Health Research England, Medical Research Council, NHS Blood and Transplant and the Rosetrees Trust.

EGCUT CD4+ and CD8+ datasets

The EGCUT CD4+ and CD8+ datasets have been described previously (Kasela et al., 2017). Briefly, the cohorts consist of healthy gene donors of the Estonian Genome Center of The University of Tartu. The study was approved by the Ethics Review Committee of Human Research of the University of Tartu, Estonia (permission no 206/T-4, date of issue 25.08.2011), and it was carried out in compliance with the Helsinki Declaration. Written informed consent to participate in the study was obtained from each individual prior to recruitment. All methods were carried out in accordance with approved guidelines. Gene expression data is available in Gene Expression Omnibus (GSE78840).

Genotype and gene expression data preprocessing, QC and analyses were performed by the eQTLGen analysis plan for Illumina arrays, in the same way as in discovery cohorts. After data processing and QC, 293 CD4+ samples and 283 CD8+ samples were added to analyses.

References

Kasela, S., Kisand, K., Tserel, L., Kaleviste, E., Remm, A., Fischer, K., ... Milani, L. (2017) Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. PLoS Genetics, 13(3): e1006643.

CD4 and CD8 meta-analyses

To maximize the replication power, the results of CD4 and CD8 replication analyses from EGCUT and Fairfax/Momozawa/Rotival were meta-analyzed by weighted Z-score method, using the same approach used in the discovery meta-analysis. For that, summary association statistics from Fairfax/Momozawa/Rotival replication analyses were converted to Z-scores by formula:

$Z = \beta / se(\beta)$,

where β is slope from the association test and $se(\beta)$ is the standard error of β. Multiple testing correction was performed by Benjamini-Hochberg method.

GTEx tissues

cis-eQTL replication

We collected genome-wide *cis*-eQTL summary stats for all tissues tested in the GTEx cohort (v7) from the GTEx web portal (gtexportal.org). Variants tested in our analysis were then matched to variants tested in GTEx by genomic position (both b37), while genes were matched on Ensembl (ENSG) gene id. We then converted the reported slope and standard error estimates to z-scores by dividing them by each other. To determine significance, we acquired the file containing significant eGenes from the GTEx web portal and used the 'pval_nominal_threshold' column to determine the P-value significance threshold per gene.

trans-eQTL replication

We collected fully processed, filtered and normalized gene expression matrices and covariates for each tissue used in GTEx eQTL analysis (v6p) (GTEx Consortium, 2017) from the GTEx web portal (gtexportal.org). For each eQTL identified in eQTLGen, we tested if the corresponding SNP and the gene are associated in each GTEx tissue using matrix-eQTL (Shabalin et al., 2012) and controlling for the covariates used by the GTEx project (three genotype PCs, genotyping platform, sex, and PEER factors estimated from expression data). To get an estimate of the null distribution for each tissue, we also made the same set of tests for 10 permutation rounds, permuting the genotype labels in order to break any true link between genotype and expression.

eQTS replication

For eQTS replication in GTEx v7, we retrieved genotypes derived from Whole Genome Sequencing (WGS) from dbGAP (www.ncbi.nlm.nih.gov/gap). Genotype data was then harmonized with Genotype Harmonizer using the same procedure as those in our discovery datasets. Briefly, variants with call-rate > 95%, minor allele frequency > 1% and Hardy-Weinberg P-value > 0.0001 were matched against the GIANT phase 1v3 release of 1000 genomes, and variant IDs were updated to match the reference. Using the harmonized genotypes, polygenic

scores were calculated in the same manner as for our discovery datasets. RNA-seq based gene expression levels were retrieved from the GTEx web portal (gtexportal.org) for each tissue. Since polygenic scores can be highly population specific, we first removed all non-European individuals and consequently corrected the gene expression levels for the covariates for each tissue (three genotype PCs, genotyping platform, sex, and PEER factors estimated from expression data). To minimize the contribution of *cis*-eQTLs to the eQTS replication signal, we first performed an iterative conditional *cis*-eQTL analysis on the residual gene expression levels. In this analysis, we iteratively regressed out the strongest significant *cis*-eQTL per gene (FDR<0.05, using 10 permutations), until no significant *cis*-eQTL genes were identified. For the final eQTS analysis in GTEx, we calculated associations between the polygenic scores and the residual gene expression levels after correcting for these *cis*-eQTLs. For comparison, we also conducted the same analysis without removing non-European individuals.

References

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. Nature 550: 204–213.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28: 1353–8.

Acknowledgements

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from: dbGaP accession number phs000424.v7.p2 on 09/27/2018.

DNA Methylation

Whole blood DNA methylation data was acquired from the BIOS Consortium, which has been described above. In total, DNA methylation data was available for 3,814 individuals, including 188 from CODAM, 751 from LLD, 791 from LLS, 762 from RS and 1,322 from NTR. RNA-seq gene expression levels were also available for a total of 2,905 individuals, including 186 from CODAM, 732 from LLD, 692 from LLS, 741 from RS and 554 from NTR. Methylation data was acquired using the Illumina Human Methylation 450K microarray. DNA preparation protocols, including bisulfite conversion and microarray hybridization, have been described in detail previously (Bonder et al., 2016).

eQTMs, or expression quantitative trait methylation (i.e. correlation between gene expression levels and methylation levels), are required to replicate eQTL using methylation QTL (meQTL),

for two reasons. Firstly, the link between genes and methylations sites is not always known; there may be many different methylation sites surrounding a gene that may or may not be affecting gene expression levels. Secondly, to compare the direction of effect between meQTL and eQTL, the direction of the correlation of the eQTM has to be taken into account. For example, with a negatively correlating eQTM, and a positive eQTL direction of effect, we would expect the direction of effect for the meQTL to be negative as well.

The power to detect eQTMs can be improved by accounting for cis-regulatory variation explained by genetics. For the gene-expression levels, we used the RNA-seq data corrected for *cis*-eQTLs prior to *trans*-eQTL mapping. Similarly, we corrected the methylation data for previously identified *cis*-meQTLs (Bonder et al., 2016). Since a single methylation site can be independently associated with multiple SNPs, we also corrected for independent SNP associations, which included up to 11 independent effects per methylation site (Bonder et al., 2016). In the case of multiple independent associations, PCA was used to orthogonalize the genotype data, and the resulting PCs were used to subtract the genotype effect. Finally, we used Spearman's ranked correlation to correlate the corrected methylation levels with the corrected gene expression levels and limiting the distance between methylation probe and gene midpoint to 1 megabases. To determine significance, we used FDR<0.05 as determined using 10 permuted datasets, applying the same framework as for our eQTL analysis. In total, we identified 57,786 significant eQTM, representing 9,675 genes.

To perform replication of *trans*-eQTL, we selected the methylation probe from the eQTM analysis that showed the strongest significant association with the *trans*-eQTL gene, and calculated the association between that probe and the *trans*-eQTL SNP. Consequently, we generated 40,590 unique SNP-methylation probe pairs, of which we were able to test 38,528 (*trans*-meQTL) resulting in 1,320 significant effects (FDR<0.05; 10 permutations).

References

Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

Single cell RNA-seq datasets

OneK1K

Ethics approval

The study was approved by the Tasmanian Health and Medical Human Research Ethics Committee (H0012902). Informed consent was obtained from all participants.

Isolation and preparation of PMBCs

Peripheral blood samples were collected from 1,083 individuals into vacutainer tubes containing either FICOLL™ and sodium heparin (8mL CPT™; BD Australia, North Ryde, NSW; 362753) or K2EDTA (10mL; BD Australia, North Ryde, NSW; Catalogue: 366643). Within 4 hrs of collection, peripheral mononuclear cells (PBMCs) were isolated from the CPT tubes according to the manufacturer's instructions. All PBMCs were kept in chilled fetal Bovine Serum (F9665; Sigma-Aldrich), RPMI-1640 medium (R8758; Sigma-Aldrich) and dimethyl sulfoxide (472301; Sigma-Aldrich). Isolated PBMCs were cryopreserved using RPMI-1640 medium including 40% FCS, RPMI, 10%DMSO. A 1mL aliquot per sample of cryopreserved cells was thawed in a 37^{o}C water bath, washed with 9mL Iscove's Modified Dulbecco's Media (IMDM; Life Technologies; 12440061) and 5% Fetal Bovine Serum (FBS; Bovogen; SFBS-FR), and resuspended in 900uL IMDM and 10% FBS. Cells were counted using a Countess II Automated Cell Counter

(ThermoFisher; AMQAX1000) and Trypan Blue viability stain (Life Technologies; T10282), and equal numbers of live cells were combined for 12-14 samples per pool.

Single-cell library preparation and sequencing

Pooled single-cell suspensions partitioned and barcoded using the 10X Genomics Chromium Controller and the Single Cell 3' Library and Gel Bead Kit version 2 (PN-120237). The pooled cells were super-loaded onto the Chromium Single Cell Chip A (PN-120236) to target 20,000 cells per pool. Libraries for all samples were multiplexed and sequenced across five 2×150 cycle S4 flow cells on an Illumina NovaSeq 6000.

Alignment and initial processing of sequencing data

The Cell Ranger Single Cell Software Suite (version 2.2.0) was used to process data produced by the Illumina NovaSeq 6000 sequencer into transcript count tables. Raw base calls from multiple flow cells were demultiplexed into separate pools of samples. Reads from each pool were then mapped to the GRCh37/hg19 genome (release 84) using STAR.

Demultiplexing and doublet identification

Cells for each individual were identified using the Demuxlet tool (Kang et al., 2018). The most likely individual for each droplet was determined using the genotype posterior probability estimate from the imputation of 265,053 exonic SNPs (R2>0.3 and MAF>0.05). In all approaches, α was set to 0.5, assuming a 50/50 ratio and other parameters were kept as default. Droplets that were identified as doublets by both Demuxlet and additional tool, Scrublet (Wolock et al., 2019) were removed from the dataset.

Cell type classification

The distributions of the total number of UMIs, number of genes, and the percentage of mitochondrial gene expression were normalized using ordered quantile transformation in each pool. The effect of sequencing depth variation due to any technical errors was removed by applying the SCTransform method (Hafemeister et al., 2019) to the variance of the gene UMI count matrix.

Cells were classified using supervised and unsupervised approaches. In the supervised classification, a reference signature matrix was built using purified PMBC data (Zheng et al., 2017) and the cosine similarity of each cell against all reference cell types was calculated. The cells were labeled based on the shortest cosine distance to reference cell type across the layers of the hierarchy. Then a graph-based unsupervised clustering was applied at the end of every hierarchy. Any misclassified cells were relabelled using known markers.

Genotypes

Samples were imputed with the Haplotype Reference Consortium panel (HRC r1.1 2016; R2>0.8; MAF > 0.01) and then harmonised to the 1000G phase1 v3 reference panel, in line with the eQTL cookbook followed by the eQTLGen cohorts.

References

Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. BioRxiv, 576827. https://doi.org/10.1101/576827

Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat.*

Commun. **8**, 14049 (2017).

1M-scBloodNL data

Ethics approval

The LifeLines DEEP study was approved by the ethics committee of the University Medical Center Groningen, document number METC UMCG LLDEEP: M12.113965. All participants signed an informed consent form before 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.

Isolation and preparation of PMBCs

Previously published PBMC scRNA-seq data of 45 donors (Van der Wijst et al., 2018) from the general population Lifelines Deep cohort (Tigchelaar et al., 2015) was combined with newly generated PBMC scRNA-seq data of another 112 donors from the same cohort. For all donors, blood was collected into EDTA-vacutainers (BD) and within 2h PBMCs were isolated with Cell Preparation Tubes with sodium heparin (BD). Cells were cryopreserved, thawed and handled as described before (Van der Wijst et al., 2018), with the exception that the newly generated dataset contains sample pools of eight instead of six individuals. For each individual we aimed at a targeted recovery of 1,000 viable single cells.

Single-cell library preparation and sequencing

Single cell libraries were generated using the 10x Chromium controller (10x Genomics) in combination with the Single Cell 3' version 2 (117 samples: Single Cell A Chip Kit, PN-120236 and Single Cell 3′ Library & Gel Bead kit v2, PN-120237) or version 3 (40 samples: Single Cell B Chip Kit, PN-1000073 and Single Cell 3′ Library & Gel Bead kit v3, PN-1000075) reagents according to the company's instructions (document CG00052 and CG000183). These libraries were multiplexed and sequenced on an Illumina NovaSeq 6000 using a 150bp paired-end kit, per BGI (Hong Kong) sequencing guidelines. In total, we captured 130,681 cells in this new dataset that were sequenced to an average depth of 45k.

Alignment and initial processing of sequencing data

The Cell Ranger Single Cell Software Suite (version 3.0.2) was used to process data produced by the Illumina NovaSeq 6000 sequencer into transcript count tables. Raw base calls from multiple flow cells were demultiplexed into separate pools of samples. Reads from each pool were then mapped to the GRCh37/hg19 Cell Ranger 3.0.0 reference genome using the STAR implementation within Cell Ranger.

Demultiplexing and doublet identification

Cells for each individual were identified using the Demuxlet tool (Kang et al., 2018). Exonic SNPs were filtered for SNPs with MAF>0.02 and were then used as input for Demuxlet. Droplets where the doublet likelihood score minus the singlet likelihood score was less than 0.25 were removed. An additional filter was applied to remove cells with a singlet likelihood score between 0 and 25, with more than 2,000 expressed genes. These cells were excluded for any further analysis.

Cell type classification

Cells where mitochondrial gene content was over 8% or 15% were removed for cells sequenced with the V2 chemistry or V3 chemistry, respectively. Furthermore, cells in which fewer than 200 genes were expressed were removed for analysis. Expression per cell was log transformed and

scaled to 10,000 reads. Technical errors were tackled using SCTransform (Hafemeister et al., 2019) with default settings and regressing out mitochondrial gene content.

Cells were clustered using Seurat's FindClusters algorithm using a resolution of 1, and either 30 or 20 principal components for the V2 and V3 chemistry respectively, based on the PC elbow plot. After clusters were identified, cell types were assigned a cell type based on differential marker expression found using the FindMarkers function in Seurat and known cell type marker genes.

Genotypes

Genotypes were processed as in Van der Wijst et al. 2018 as part of the Lifelines Deep consortium. Genotypes were phased using Eagle v2.330 (Loh, et al. 2016) and imputed with the HRC reference panel (McCarthy, et al. 2016), using the Michigan imputation Server (Das, et al. 2016).

References

Das, S. et al. Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287 (2016).

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. BioRxiv, 576827. https://doi.org/10.1101/576827

Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48, 1443–1448 (2016).

McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283 (2016).

Tigchelaar, E. F., Zhernakova, A., Dekens, J. A. M., Hermes, G., Baranska, A., Mujagic, Z., … Feskens, E. J. M. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*, *5*(8), e006772. http://doi.org/10.1136/bmjopen-2014-006772

Van Der Wijst, M. G. P., De Vries, D. H., Brugge, H., Westra, H. J., & Franke, L. (2018). An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine*, *10*(1), 96. https://doi.org/10.1186/s13073-018-0608-4

Single cell replication meta-analysis

The data from the OneK1K cohort and 1M-scBloodNL were meta-analyzed in an effort to replicate the eQTLGen *trans*-eQTLs. 982 samples from the OneK1K cohort, and respectively 45, 72 and 40 samples from the subgroups of the 1M-scBloodNL cohort were included in the analyses.

Where cell types were identified with more resolution in one of the cohorts, the expression matrices were added to attain a consensus for the meta-analysis:

OneK1K	**1M-scBloodNL**

CD4+ KLRB1- T cell	CD4+
CD4+ KLRB1+ T cell	
CD4+ SOX4+ T cell	
CD8+ GNLY+ NKG7+ T cell	CD8+
CD8+ LTB+ T cell	
CD8+ S100B+ T cell	
XCL1- NK	NK
XCL1+ NK	
IgJ+ B cell	Plasma
TCL1A- FCER2- B cell	B cells
TCL1A+ FCER2+ B cell	
Monocyte CD14+	Classical Monocytes
Monocyte FCGR3A+	Non-classical monocytes
Dendritic cell	Plasmacytoid DC
	Myeloid DC

We tested SNP-gene combinations that were significant in the *trans*-eQTL analysis and where the gene was sufficiently expressed (a missing sample fraction of <20%). We used the eQTL mapping pipeline for these analyses, with the same settings as the main analyses.

GWAS summary statistics

Full association summary statistics were downloaded from several publicly available resources and are indicated in **Supplementary Table 13**. Studies done exclusively in non-European cohorts were omitted. Filters applied to the separate data sources are indicated below. All the dbSNP rs numbers were standardized to match GIANT 1000G p1v3 and the directions of the effects were standardized to correspond GIANT 1000G p1v3 minor allele. SNPs with different opposite-strand alleles compared to GIANT alleles were flipped. SNPs with A/T and C/G SNPs as well as SNPs with different alleles GIANT 1000G p1v3 (tri-allelic SNPs, indels, unknown alleles) were removed from the analysis. Genomic control was applied to all the P-values for the datasets that were not genotyped by Immunochip, Metabochip or Exome chip. Additionally, genomic control was skipped for datasets that did not have all associations available[1] and all the datasets from GIANT consortium, as these had the application of genomic control indicated on the web site. In all, 1,263 summary statistic files were added to the analysis.

AdipoGen

GWAS meta-analysis summary statistics for adiponectin levels[2] were downloaded from (http://www.mcgill.ca/genepi/adipogen-consortium). Data filtering involved the removal of SNPs for which summary statistics were unavailable and SNPs that were tested in less than half of the maximal number of samples in meta-analysis.

CARDIOGRAM

Data on coronary artery disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and CARDIoGRAM Exome investigators, and have been downloaded from www.CARDIOGRAMPLUSC4D.org. For two studies[3,4], SNPs that were not present in more than half of the samples or had Cochran's Q test $P \leq 0.0001$ were removed. For the CARDIoGRAMplusC4D Consortium (2015) study only the threshold for Cochran's Q test $P > 0.0001$ was applied.

CHARGE

GWAS summary statistics for several fatty acids[5-8] were collected from consortium web site (http://www.chargeconsortium.com/main/results). SNPs were filtered based on heterogeneity test P-value ≥ 0.0001 and presence in more than half of the cohorts.

CKDGen

GWAS meta-analysis summary statistics for kidney-related functions[9,10] were collected from http://fox.nhlbi.nih.gov/CKDGen/. SNPs not present in at least half of the maximum number of samples in meta-analysis were filtered out.

CONVERGE

GWAS summary statistics for major depression[11] were downloaded from https://www.med.unc.edu/pgc/files/resultfiles/. No filters were available and applied.

DIAGRAM

GWAS meta-analysis summary statistics for type 2 diabetes[12,13] were collected from http://diagram-consortium.org/downloads.html. We removed SNPs that were not present in more than half of the samples in the meta-analysis.

EAGLE

GWAS meta-analysis summary statistics for eczema[14] and preschool internalizing problems[15] were collected from https://data.bris.ac.uk/data/dataset/28uchsdpmub118uex26ylacqm and http://www.tweelingenregister.org/fileadmin/user_upload/EAGLE/Internalizing.zip. SNPs not present for at least half of the maximum number of European cohorts or having Cochran's Q test $P \leq 0.0001$ were filtered out.

EGG

GWAS summary statistics for childhood growth phenotypes[16-23] were collected from http://egg-consortium.org/. If sample sizes were available in the summary statistics files, we included SNPs which were tested in more than half of the samples.

GABRIEL

GWAS meta-analysis summary statistics for asthma[24] were collected from http://www.cng.fr/gabriel/results.html. SNPs not present in at least half of the maximum number of samples in meta-analysis or having Cochran's Q test $P \leq 0.0001$ were filtered out.

GEFOS

Whole-genome sequencing, whole-exome sequencing, and deep imputation of genotype data based meta-analysis summary statistics for bone density traits from Zheng et al[25] were collected from http://www.gefos.org/?q=content/data-release-2015. SNPs not tested in at least half of cohorts in meta-analysis (3 out of 5) were filtered out.

Acknowledgements

We thank GEFOS-seq consortia for making these data available for research use.

GIANT

GWAS meta-analyses summary statistics for BMI[26,27], hip and waist circumference[28], and height[29] were downloaded from http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files. All SNPs were filtered based on their presence in more than half of the samples in meta-analysis or, if available, presence in more than half of the cases and controls in the meta-analysis.

GLGC

GWAS meta-analyses summary statistics for lipid traits[30] were collected from http://csg.sph.umich.edu//abecasis/public/lipids2013/. As results were already sample-size filtered (N>50,000 in GWAS and N>20,000 in Metabochip datasets), we did not apply any additional filters to these data. Additional summary statistics[31] were collected from http://csg.sph.umich.edu/abecasis/public/lipids2010/. In this dataset, we included only SNPs that were present in at least half of the cohorts included in the meta-analysis.

GPC (The Genetics of Personality Consortium)

GWAS meta-analyses summary statistics for several personality traits[32-34] were collected from http://www.tweelingenregister.org/GPC/. All SNPs not tested in more than half of the samples were filtered out.

GUGC

GWAS meta-analysis summary statistics for serum urate and gout[35] were collected from http://metabolomics.helmholtz-muenchen.de/gugc/. Data was not filtered as all reported SNPs were already present in at least 75% of all the samples in meta-analysis. P-value after genomic control was used.

HemGen

GWAS meta-analyses for red blood cell traits[36] summary statistics were downloaded from European Genome-phenome Archive study accession EGAS00000000132 (public access dataset: access provided by EGA HelpDesk). SNPs were filtered based on the presence in at least half of the studies in the meta-analysis.

HRgene consortium

GWAS summary statistics for heart rate[37], fat percentage[38] and leptin[39] were downloaded from the website of the HRgene consortium/Loos lab (https://walker05.u.hpc.mssm.edu/). SNPs were filtered based on presence in at least half of the samples in the meta-analysis and a heterogeneity test P-value ≥0.0001 (if reported).

IGAP

GWAS meta-analysis summary statistics for Alzheimer's disease[40] were collected from http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php.

Data was not filtered as sample-size and heterogeneity information was not available, and meta-analysis consisted of only the SNPs that were genotyped or imputed in at least 40% of both, cases and controls.

Acknowledgements

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data, but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i–Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728.

IIBDGC (International Inflammatory Bowel Disease Genetics Consortium)

GWAS meta-analysis summary statistics for European cohorts[41] were downloaded for IBD, ulcerative colitis and Crohn's disease (http://www.ibdgenetics.org/downloads.html). All files were filtered to include SNPs tested in at least half of the datasets in meta-analysis and having heterogeneity test P-value > 0.0001.

Immunobase

GWAS summary statistics for several autoimmune diseases were collected from Immunobase (http://www.immunobase.org/; accessed 26 April 2016). The following traits were downloaded: ulcerative colitis[42], multiple sclerosis[43,44], systemic lupus erythematosus[45], primary biliary cirrhosis[46,47], celiac disease[48,49], narcolepsy[50], juvenile idiopathic arthritis[51], rheumatoid arthritis[52,53], type 1 diabetes[54] and psoriasis[55].

MAGIC

Several GWAS meta-analysis summary statistics for glucose and insulin traits[56-63] have been contributed by MAGIC investigators and have been downloaded from http://www.magicinvestigators.org. No additional filters were available and applied to these data.

Acknowledgements

Data on glycemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

MAGNETIC

GWAS summary statistics for metabolites from[64] were downloaded from http://www.computationalmedicine.fi/data. SNPs were filtered by the presence in at least half of the samples in the meta-analysis.

Metabolomics

Metabolomics GWAS summary statistics[65] were downloaded from http://mips.helmholtz-muenchen.de/proj/GWAS/gwas/gwas_server/shin_et_al.associations.tar.gz. As this study was a meta-analysis incorporating only two cohorts, we did not apply any filters based on number of cohorts in the meta-analysis or heterogeneity test P-value.

PGC (Psychiatric Genomics Consortium)

Publicly available GWAS meta-analyses summary statistics for several personality and cognitive traits[66-72] were downloaded from https://www.med.unc.edu/pgc/results-and-downloads.

Interim results for autism spectrum disorder are cited as follows:

Autism Spectrum Disorder Working Group of the Psychiatry Genomics Consortium. Dataset: PGC-ASD summary statistics from a meta-analysis of 5,305 ASD-diagnosed cases and 5,305 pseudocontrols of European descent (based on similarity to CEPH reference genotypes) (March 2015). (available at: http://www.med.unc.edu/pgc/results-anddownloads)

No additional filters were applied to these datasets.

Project MinE

GWAS summary statistics for ALS risk from[73] were downloaded from http://databrowser.projectmine.com. Summary statistics from mixed linear model analysis were used, and no additional filters were available.

Acknowledgements

We would like to thank the Project MinE GWAS Consortium for providing the data.

ReproGen

GWAS meta-analysis summary statistics for age of menopause[74] and age of menarche[75] were collected from (http://www.reprogen.org/data_download.html). No additional filters were available and applied to these data.

SSGAC

GWAS meta-analysis summary statistics for several social science outcomes[76-80] were collected from http://www.thessgac.org/#!data/kuzq8. No additional filters were available or applied to these data.

TAG

GWAS meta-analysis summary statistics on smoking behavior[81] were collected from https://www.med.unc.edu/pgc/results-and-downloads. No additional filters were available or applied to these data.

Gieger et al., 2011

GWAS summary statistics for platelet phenotypes[82] were acquired through communication with Prof. Soranzo. We excluded SNPs that were tested in less than half of the individuals in the analysis.

Acknowledgements

We thank Prof. Nicole Soranzo for her help in providing these summary statistics.

Hyde et al., 2016

GWAS summary statistics for major depressive disorder were curated from the supplementary material of original publication[1]. This data involved only the 10,000 most significant SNPs ($P < 5.3\times10^{-5}$).

Orrù et al., 2013

GWAS summary statistics for immune cell traits from the paper[83] were acquired from http://www.irgb.cnr.it/facsdataexplorer. No additional filters were applied to these data.

Acknowledgements

We thank Mr. Luiz Fernando Pereira for his help in downloading these summary statistics.

Roederer et al., 2015

GWAS summary statistics from immune cell traits[84] were downloaded from ftp://twinr-ftp.kcl.ac.uk/ImmuneCellScience/2-GWASResults/. The same filters were applied as in the original study: Hardy-Weinberg disequilibrium test P-value < 0.0001, MAF>0.05 and call rate > 95%.

Full summary statistics datasets from GRASP website

Exome-chip-based full summary statistics for platelet traits[85], white blood cell counts[86] and red blood cell counts[87] were downloaded from the GRASP 2.0 website (https://grasp.nhlbi.nih.gov/FullResults.aspx)[88]. No additional SNP filters were available and applied.

Supplementary Methods

Conditional *cis*-eQTL analyses

Summary statistics based conditional analysis

For the 14,115 samples profiled with Illumina expression arrays, we applied summary-statistics-based conditional analysis on meta-analyzed Z-scores. The analysis strategy was modified from[89] performing joint effect estimation from equation (5) and conditional analysis from equation (19) of the aforementioned paper.

We used standardized Z-scores re-calculated using only cohorts of European descent and the formula: $Z / \sqrt{n}$.

For each Illumina probe and chromosome, the following steps were taken:

1. Reference genotype data from 1kG EUR was used. After standardizing the genotypes to mean zero and variance 1, correlation matrix *C* was calculated. To avoid overestimation of the correlations, we used a shrinking parameter of 1 - λ, with $\lambda = 2 / \sqrt{n}$ and *n* being the sample size in the reference panel. If alleles did not match, the sign of Z was swapped.
2. We removed SNPs for which $N < (N_{total} \times 0.9)$, SNPs that were missing from reference panel and SNPs for which the study-specific allele frequency deviated from reference panel for more than 0.05.
3. We selected the top SNP (denoted as *B1*), by choosing the SNP with the largest absolute standardized effect size.
4. We conditioned all remaining SNPs on this top SNP(s) (denoted as *B2*): $B2_{cond} = B2 - B1\, C_{11}^{-1} C_{12}$
5. We removed B2 SNPs that have $|C_{12}| > 0.9$
6. We ranked the B2 SNPs according to decreasing absolute $B2_{cond}$ and select top SNP. Include this SNP to B1 SNPs.
7. We calculated joint effect sizes: $B1_{joint} = B1\, C_{11}^{-1}$. Only keeping B1 SNPs that had $(|B1_{joint}| / \times \sqrt{n}) > 4$, *n* being the sample size in the reference panel.
8. We repeated steps 3 to 7 until no additional SNPs were identified to have |Z| > 0.4 (lowest *Z* corresponding to FDR < 0.05).
9. We calculated joint effect sizes for B1 SNPs, then converted the joint effect sizes to joint Z statistics.

Iterative conditional analysis

For individual datasets profiled by RNA-seq and Affymetrix arrays, we applied iterative conditional *cis*-eQTL mapping. First, *cis*-eQTL mapping was conducted using the same settings as the discovery *cis*-eQTL mapping analysis. For all significant (FDR<0.05) *cis*-eQTL effects, the most significant SNP effect for each gene was regressed out from the gene expression matrix. The next round of *cis*-eQTL mapping analysis was conducted on the adjusted expression matrix while

testing only genes that had any significant *cis*-eQTL effect prior regression. The analysis was performed iteratively until no significant (FDR<0.05) effects remained for a given gene.

Sample size estimation for Framingham Heart Study

Being a family-based cohort, the Framingham Heart Study was the only dataset in the analysis that did not consist of unrelated individuals. Although the analysis strategy took family relationship into account, we needed to determine the effective sample size to use proper weight for this dataset in the weighted Z-score meta-analysis. To do so, we selected 20 random genes and used eQTL effects for genome-wide SNPs to estimate effective sample size (N_{eff}), based on the formula:

$$N_{eff} = \frac{\sum_{i=1}^{m}[var(y) - 2p_i(1-p_i)\beta_i^2]/[2p_i(1-p_i)S_i^2]}{m}$$

Where $var(y)$ is the variance of the phenotype, p is the minor allele frequency of the SNP, β is the estimated SNP effect on the expression level of the gene, *S* is the corresponding standard error of β, and m is the number of genome-wide SNPs.

As the effective sample size was not dramatically different when compared to the real one (mean N_{eff} over all 20 genes was 4,793; median N_{eff} = 4,844, as compared to real N = 5,075; **Supplementary Table 19**) and thus will not influence the results of the combined meta-analysis considerably, we opted to use the real sample size as a weight in the meta-analysis.

Removal of cross-mapping artefacts from *trans*-eQTL results

When a *trans*-eQTL gene has similar paralogous genes in close proximity (< 5Mb) of a given eQTL SNP, the apparent *trans*-eQTL effect may actually reflect a much stronger *cis*-eQTL effect. This might be caused by non-unique binding of the array probe (probe binds with several similar mRNA molecules) or some non-unique read mapping in case of RNA-seq (read is assigned to several similar regions in the genome). While both scenarios should be corrected by filtering out multi-mapping Illumina array probes and by not counting the RNA-seq reads assigned to multiple genomic features, there might still be some non-detected cases.

To remove such false positive *trans*-eQTLs, we created sets of 35 bp "reads" from the human reference genome (ENSEMBL v71) for each significant (FDR<0.05) *trans*-eQTL gene. To span the gene sequence, we used a shifting window approach, with each consecutive window shifting 2bp, while also generating reads spanning exon-exon boundaries. We created 10Mb sequences centered around each significant *trans*-eQTL SNP. Then, we mapped the reads generated for the gene to the 10Mb SNP region using BWA-mem v0.7.15[90] for each *trans*-eQTL SNP-gene pair. Finally, for each *trans*-eQTL, we summed the fractions of aligned base pairs over all the reads generated for given gene and divided this by the number of reads for a given gene, resulting in a proportion of the gene mapped within 5Mb of the SNP.

As *trans*-eQTLs with high proportions of genes mapping within the SNP region are more likely to actually be *cis*-eQTL effects, we reasoned that those should show nominal replication (uncorrected P<0.05) in many GTEx v6p tissues. Based on visual inspection of diagnostic plot (**Supplementary Figure 14A**), we opted to use 5% of gene mapping to the vicinity of a SNP as a threshold to declare a *trans*-eQTL to be potentially caused by cross-mapping. This strategy flagged 8,984 (12.2%) out of 73,298 significant *trans*-eQTLs as potentially cross-mapping. After re-calculating the FDR after SNP pruning, 59,786 (81.6%) effects remained significant (FDR<0.05).

26 *trans*-eQTLs showing very low cross-mapping (<5%), yet high nominal replication rate in GTEx tissues (>30% of tested tissues with P<0.05, >10 tissues tested for each *trans*-eQTL), were further investigated by forest plots (**Supplementary Figure 14B**). 8 out of the 26 *trans*-eQTLs showed strong and unidirectional effects for the majority of RNA-seq and array platforms, suggesting that these are not caused by cross-mapping. The remainder of the *trans*-eQTLs were tested mainly or exclusively in RNA-seq datasets. However, investigation of regions near the eSNP in the UCSC browser (https://genome.ucsc.edu/, most up-to-date hg38 build) did not identify any genomic features that are likely to be paralogous with its *trans*-eQTL gene (e.g. pseudogenes of the *trans*-eQTL gene). Likely cross-mapping *trans*-eQTLs had generally higher effect sizes in most different cell-type-specific datasets (**Supplementary Figure 14C**) as well as individual RNA-seq-based LCL replication datasets, but not in Illumina array dataset ALSPAC (**Supplementary Figure 14D**). Considering that the effect sizes of *cis*-eQTLs are generally stronger than *trans*-eQTLs, these data suggests that we have removed the large majority of *trans*-eQTL effects which were actually caused by cross-mapping with *cis*-eQTL loci.

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 1000G p1v3 EUR (**Supplementary Figure 15**), and the effect direction of each dataset was compared against the meta-analyzed effect (**Supplementary Figure 16A-C**).

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 **Supplementary Figure 17**. We used the discovery *trans*-eQTL analysis results as an input, confined ourselves to those effects that were present in the datasets we had direct access to (BBMRI-BIOS+EGCUT; N=4,339), and showed nominal $P<8.3\times10^{-06}$ 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 (FDR<0.05), we defined the locus by adding a ±1 Mb window around it. Next, for each *trans*-eQTL gene, we ran iterative conditional *trans*-eQTL analysis using all loci for a given *trans*-eQTL gene. We then evaluated the LD between all conditional lead *trans*-eQTL SNPs and lead *cis*-eQTL SNPs using a 1 Mb window and $R^2>0.8$ (1kG p1v3 EUR) as a threshold for LD overlap.

cis-eQTL - *trans*-eQTL interaction analyses

We aimed to identify local *cis*-eQTL genes that affect the *trans*-eQTL effect by changing its strength or direction and might therefore serve as potential mediators. We used a G × E interaction model to test this:

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

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. We omitted *trans*-eQTL SNP locating to HLA region from those analyses because of the complex structure of this region. On top of the gene expression normalization that we used for discovery analyses, we used

a rank-based inverse normal transformation to enforce a normal distribution before fitting the linear model. This is identical to the normalization used by Zhernakova et al[91] in their G × E interaction eQTL analyses. We fitted this model separately to each of the cohorts that are part of the BIOS consortium and to EGCUT. We transformed the interaction P-values to Z-scores and used the weighted Z-score method[92] to perform a meta-analysis of 4,339 samples. The Benjamini-Hochberg procedure[93] was used to limit the FDR to 0.05. The plots in **Supplementary Figure 18** were created with the default normalization, and the regression lines are the best-fitting lines between the mediator gene and the *trans*-eQTL gene, stratified by genotype.

Cell-type-composition effects of *trans*-eQTLs and eQTS

Dataset

We used data from a subset 3,831 BIOS individuals to which we had direct access. We further narrowed our sample set down to 1,858 individuals for whom the measured cell metric data was available for at least ⅔ of measured cell metrics. All samples were part of discovery meta-analyses.

Measured cell metrics

Several cell types were counted in peripheral blood from each of the BIOS cohort participants, but cohorts differed in the availability. Cells were counted as an absolute number in a liter of blood (white blood cell count, red blood cell count, platelet count), or as a percentage of the white blood cell count (neutrophil percentage, lymphocyte percentage, etc.). Out of 24 cell metrics, we excluded eight (LUC, LUC%, RBC, RDW, MCH, MPV, MCHC, MCV) because these measurements were not available for the large fraction of samples, hindering the estimation of the combined effect of measured cell metrics on *trans*-eQTLs and eQTS. All measured cell metrics are summarized in **Supplementary Table 20**.

Estimated cell counts

We estimated the cell counts of 33 different cell types using Decon-cell, part of the Decon2 method[94]. Decon-cell was trained using information from the independent 500FG cohort, which includes detailed cell type measures as well as RNA-seq expression profiles[95]. Next, the prediction model was used to impute cell proportions based on the BIOS gene expression matrix. Predicted cell metrics are summarized in **Supplementary Table 20**.

Cell type interaction analyses

Here we used data from a subset of up to 1,858 BIOS Consortium samples for which 49 measured and predicted cell type metrics were available. For these analyses we tested only effects where the SNP had a MAF>0.05 in each BIOS cohort.

All 49 cell metrics were transformed by inverse normal transformation prior to analyses. For gene expression, we used the same preprocessing as in the discovery meta-analyses, including correction for expression PCs and regression of *cis*-eQTL effects. In addition to the standard preprocessing, the expression of each gene was transformed using inverse normal transformation.

For multivariate linear models, analyses were conducted using R v3.4.4, data.table v1.12, tidyverse v1.2.1, broom v0.5.1 and the pheatmap v1.0.12 packages. For each BIOS cohort, linear models were fitted for each *trans*-eQTL identified in meta-analysis (FDR<0.05), using lm() function from R. For eQTS analyses, PGS was used instead of SNP.

Three different interaction models were fitted for each *trans*-eQTL and eQTS:

$t = \beta_0 + \beta_1 \times s$

$t = \beta_0 + \beta_1 \times c_1 + \beta_2 \times c_2 + \ldots + \beta_{49} \times c_{49} + \beta_{50} \times s$

$t = \beta_0 + \beta_1 \times c_1 + \beta_2 \times c_2 + \ldots + \beta_{49} \times c_{49} + \beta_{50} \times c1 \times s + \ldots + \beta_{99} \times c_{49} \times s + \beta_{100} \times s$

where t is the expression of *trans*-eQTL/eQTS gene, c is cell-type metric, and s is a dosage of *trans*-eQTL SNP or scaled value of polygenic score. P-values from each term of the linear model (main effects and interaction effects) were converted to signed Z-scores and effects were meta-analyzed by weighted Z-score method, using the square root of per-cohort sample size as weight.

To determine the effect of cell-type composition on *trans*-eQTLs, we applied models and assessed the SNP main effect. Here we used the same significance thresholds as determined by the permutation-based FDR in the discovery meta-analyses.

To determine the likely cell types where *trans*-eQTLs or eQTSs can manifest, we applied the third model with the difference that no PCs were removed from gene expression data prior to analysis and queried the individual interaction term for each cell metric. A Benjamini-Hochberg FDR[93] across all interaction P-values was used to determine significance in this analysis.

Identifying cell types correlating with polygenic scores

To identify the cell types that are most affected by genes associated with the polygenic scores for autoimmune diseases (**Supplementary Figure 13D**), we investigated scRNA-seq data[96]. Since each disease is represented using multiple GWAS P-value cutoffs, we first opted to calculate an average Z-score per gene, averaging over the P=0.01, P=0.001, $P=1\times10^{-4}$, $P=1\times10^{-5}$, and $P=5\times10^{-8}$ thresholds. For this purpose, we made use of the non-PC-corrected eQTS results, since this dataset is likely to still contain major effects of cell-type-composition differences. From the resulting Z-score vector, approximately 5,000 significant genes were selected, as determined from the average Z-scores. We then calculated correlations between the Z-scores for the selected genes and gene expression levels from a scRNA-seq dataset consisting of ~25,000 peripheral blood mononuclear cells isolated from 45 individuals[96]. We then repeated this analysis in the 10 permuted eQTS datasets, and calculated an empirical FDR threshold that was used to investigate individual traits. From this analysis we observed that eQTS signatures for systemic lupus erythematosus, ulcerative colitis and celiac disease showed significantly (FDR<0.05) increased or decreased expression in at least 10 cells.

Enrichment analyses

Overview of enrichment analyses

The table below includes enrichment analyses that were performed in the study, and describes for each analysis the method, the tested annotations, the test set, the background set and the null hypothesis.

Enrichment analysis	Method	Tested annotations	Test set	Background	Null hypothesis
TF enrichment analysis for rs17087335 *trans*-eQTL genes.	One-sided Fisher's exact tests as implemented in GeneOverlap.	TF targets as determined by CHiP-X. Downloaded from Enrichr web site.	*trans*-eQTL genes.	All 19,942 genes tested in *trans*-eQTL meta-analysis.	There is no overrepresentation of TF targets among *trans*-eQTL genes for rs17087335.
TF and miRNA target site enrichment analyses for *hub trans*-eQTL SNPs.	One-sided Fisher's exact tests as implemented in GeneOverlap.	TF targets as determined by CHiP-X, putative TF targets based on TRANSFAC and JASPAR PWMs. miRNA targets from TargetScan and MirTarBase. Downloaded from Enrichr web site.	*trans*-eQTL genes.	All 19,942 genes tested in *trans*-eQTL meta-analysis.	There is no overrepresentation of TF or miRNA targets.
Gene Ontology (GO) enrichment for co-localizing *cis*-eQTL and *trans*-eQTL effects.	One-sided Fisher's exact tests as implemented in GeneOverlap.	GO gene sets (2018. year version). Downloaded from Enrichr web site.	*cis*-eQTL genes, showing co-localization (R^2>0.8) with any *trans*-eQTL gene.	All 16,987 genes significant in *cis*-eQTL meta-analysis.	There is no over-representation of GO annotation for *trans*-eQTL genes co-localizing with *cis*-eQTL genes.
Gene Ontology (GO) enrichment for per-phenotype *trans*-eQTL genes.	One-sided Fisher's exact tests as implemented in GeneOverlap.	GO gene sets (2018. year version). Downloaded from Enrichr web site.	*trans*-eQTL genes, stratified by GWAS phenotype.	All 19,942 genes tested in *trans*-eQTL meta-analysis.	There is no overrepresentation of GO terms for any phenotype.
Gene Ontology (GO) enrichment for per-phenotype eQTS genes.	One-sided Fisher's exact tests as implemented in GeneOverlap.	GO gene sets (2018. year version). Downloaded from Enrichr web site.	eQTS genes, stratified by GWAS phenotype.	All 19,942 genes tested in eQTS meta-analysis.	There is no overrepresentation of GO terms for any phenotype.
Gene Ontology (GO) enrichment for	One-sided Fisher's exact tests as	GO gene sets (2018. year version).	eQTS genes.	All 19,942 genes tested in eQTS meta-	There is no overrepresentation of GO

all eQTS genes.	implemented in GeneOverlap.	Downloaded from Enrichr web site.		analysis.	terms for eQTS genes.
TF enrichment for all eQTS genes.	One-sided Fisher's exact test as implemented in GeneOverlap	TF information from FANTOM5 (Abugessaisa *et al.*, 2016)	eQTS genes.	eQTS genes and TF targets.	There is no overrepresentation of TFs among eQTS genes.
TF enrichment of *cis-trans* gene pairs.	Two-sided Fisher's exact test.	TF information from Regulatory Circuits (Marbach *et al.,* 2016).	*Trans*-eQTLs converted to gene pairs.	Gene-gene combinations and TF-target pairs.	There is no enrichment of TF-target gene pairs among *trans*-eQTLs.
PPI enrichment of *cis-trans* gene pairs.	Two-sided Fisher's exact test.	InWeb protein-protein interactions (Li *et al.,* 2017).	*Trans*-eQTLs converted to gene pairs.	Gene-gene combinations and TF-target pairs.	There is no enrichment of interacting proteins among *trans*-eQTLs.
Co-regulation of *cis-trans* gene pairs.	Two-sided Fisher's exact test.	Co-regulation as derived from eQTLGen summary statistics.	*Trans*-eQTLs converted to gene pairs.	Gene-gene combinations and TF-target pairs.	There is no enrichment of Co-regulation genes among *trans*-eQTLs.
Hi-C contact between *cis-trans* gene pairs.	Two-sided Fisher's exact test.	Hi-C contacts 10kb resolution, (Rao et al, Science 2014).	*Trans*-eQTLs converted to gene pairs.	Gene-gene combinations and TF-target pairs.	There is no increased contact between the genomic locations of *trans*-eQTL genes and SNPs

TF and microRNA target enrichment analyses for hub SNPs

Hub SNPs were defined as SNPs having more than 10 *trans*-eQTL genes. TF target genes and microRNA target genes were downloaded from the Enrichr web site [97,98]. These included TF targets as assayed by ChIP-X experiments from ChEA[99] and ENCODE projects[100,101], putative TF targets based on positional weight matrices (PWMs) from TRANSFAC[102,103], JASPAR[104] and Genome Browser; predicted microRNA targets from TargetScan[105,106] and experimentally supported microRNA targets from miRTarBase[107].

A one-sided Fisher's exact test was applied to the downstream genes for all *trans*-eQTL SNPs with >10 *trans*-eQTL effects (1,050 gene sets), using the Bioconductor package GeneOverlap and 19,942 genes tested in the *trans*-eQTL analysis as a background of the analysis. Multiple testing correction was performed over all gene sets and SNPs using the Benjamini-Hochberg method. To prioritize TFs and miRNAs that are more likely to cause the *trans*-eQTL effects, we required the location of the gene encoding the TF or microRNA to be <1 Mb from the *trans*-eQTL SNP. Coordinates of genes encoding each TF were downloaded from Ensembl v75 and genomic coordinates of microRNAs were downloaded from miRBase v21[108].

Enrichment analyses for *trans*-eQTL and eQTS genes

We downloaded curated Gene Ontology gene sets (2018 version)[109] from the Enrichr web site[97]. Gene Ontology gene set over-representation analyses were conducted per each phenotype which had ≥10 downstream genes using one-sided Fisher's exact tests as implemented in GeneOverlap while using all 19,942 genes tested in the *trans*-eQTL analysis as a background of the analysis. Multiple testing correction was performed over all gene sets, using the Benjamini-Hochberg method.

For per-phenotype enrichment analyses, all *trans*-eQTL genes and eQTS genes (FDR<0.05 in discovery analysis) were stratified by corresponding GWAS phenotype prior analysis.

Separate Gene Ontology enrichment analysis was also conducted to the combined set of 2,568 significant eQTS genes (FDR<0.05 from discovery eQTS analysis), to test general enrichment for transcriptional regulation related GO terms.

Transcription factor enrichment analyses for eQTS genes

Human transcription factor annotations were downloaded from FANTOM5 SSTAR[110]. All eQTS genes (FDR<0.05 in discovery analysis) were tested for enrichment by one-sided Fisher's exact test as implemented in GeneOverlap.

PPI overlap analyses for eQTS genes

To test whether eQTS genes (FDR<0.05 from discovery analysis) had more PPI partners than the rest of the genes, we used the protein-protein interaction (PPI) data from InWeb[111]. We first intersected all eQTS genes to include only those which were available in PPI dataset. We then counted the interaction partners for each gene and compared the number of partners between eQTS genes and the rest of the genes by Wilcoxon rank sum test, as implemented in R v3.4.4.

LD overlap analyses between *cis*-eQTLs and *trans*-eQTLs

In order to evaluate the potential co-localization of association signal from two analyses which share samples, we determined the LD between the lead SNPs from both analyses. *Cis*-eQTL data from 31,684 blood samples and locus-wide conditional *trans*-eQTL analysis data from the subset of 4,339 samples (BIOS and EGCUT) were used to evaluate the potential colocalization

between *cis*-eQTL and *trans*-eQTL signals. We declared the *cis*- and *trans*-eQTL signals as co-localising when primary *cis*-eQTL and conditional *trans*-eQTL lead SNPs were in high LD (R^2>0.8; 1000G p1v3 EUR).

We downloaded curated Gene Ontology gene sets (2018 version)[109] from the Enrichr[97] web site (https://maayanlab.cloud/Enrichr/#stats). Those gene sets were used to conduct over-representation analyses (one-sided Fisher exact test) as implemented in R package GeneOverlap, while using 16,987 genes under significant *cis*-eQTL effect as a background.

Biological mechanisms explaining *trans*-eQTLs

Conversion of *trans*-eQTL results to *trans*-eQTL gene × gene P-value matrices

To better understand the biological mechanisms underlying the *trans*-eQTLs, we performed a number of enrichment analyses. We wanted to use transcription factor (TF)-target pairs, gene co-regulation, and protein-protein interaction for these enrichments. For that reason, we needed to transform the *trans*-eQTLs into a gene-by-gene matrix filled with *trans*-eQTL P-values. We converted the *trans*-eQTL results into three matrices (corresponding to the three columns in **Supplementary Figure 9**): one based on the Pascal method[112] (strategy described in **Supplementary Figure 10**), one based on *trans*-eQTL SNPs that also had *cis*-eQTL effect, and one based on the combination of these two.

Pascal method to construct gene × gene matrix

For each gene, we selected those SNPs that mapped within 50kb of the transcription start site (TSS) or transcription end site (TES). To then determine whether that set of SNPs is significantly affecting one of the 19,942 genes, we used the Pascal method to sum the effects of the different tested variants within this 50kb window while accounting for LD using the European 1000G phase 3 samples.

eQTL method to construct gene × gene matrix

To construct the gene × gene matrix while including eQTL information, we performed exactly the same steps as in the Pascal method explained above, with the exception that we only included SNPs with a *cis*-eQTL for each gene, instead of the SNPs in a 50kb window.

Combined method to construct gene × gene matrix

In the combined method we use all SNPs in a 50kb window as well as *cis*-eQTL SNPs. In our analysis, *cis*-eQTL SNPs can be up to 1 Mb away from the center of the gene. Therefore, the combined matrix includes information on more gene pairs than either the Pascal matrix or the eQTL matrix.

Gene co-regulation matrix based on eQTLGen expression data

To calculate the gene co-regulations we first needed to calculate the co-expression between genes. For that, we concatenated the 10 permuted *trans*-eQTL Z-score matrices, and filtered the list to a set of 4,586 unlinked SNPs (LD R^2<0.1), yielding a matrix of 19,942 unique genes × 45,860 independent and identically distributed variables. Since Z-scores are standardized under the null, we calculated the inner product per gene-pair in order to get the true co-expression correlation of that gene-pair. We then performed a PCA over this co-expression matrix and extracted the first 500 eigenvalues to use for the co-regulation calculation. The co-regulation of two genes was calculated by correlating the 500 eigen coefficients of the two genes, this was

done for all gene pairs to create the co-regulation matrix. This co-regulation matrix was used to perform the enrichment analyses described below.

Enriched overlap with set of transcription factor - downstream target gene-pairs

We used the above-mentioned *trans*-eQTL gene-by-gene matrices to ascertain the overlap with established sets of gene pairs. We first studied the overlap with a set of TF-downstream target gene-pairs, downloaded from RegulatoryCircuits.org[113]. We used the TF-target gene pairs predicted in lymphocytes, myeloid_leukocytes, lymphocytes_of_b_lineage, and myeloid_leukemia.

To calculate whether there is any enrichment, we compared how many of the tested *trans*-eQTL gene pairs are known TF-target pairs, and how many of the TF-target pairs are significant *trans*-eQTL gene pairs. This results in a contingency table that we use as input for a two-sided Fisher exact test to calculate the odds ratio and P-value of the enrichment. Analysis strategy is depicted in the **Supplementary Figure 10**.

Next, we expanded the TF-target matrix by including genes that were highly co-regulated with the target gene (**row 2 in Supplementary Figure 9**), genes that were highly co-regulated with the TF (**row 3 in Supplementary Figure 9**), or both (**row 4 in Supplementary Figure 9**).

Enriched overlap with co-regulated genes

We also ascertained whether the *trans*-eQTL gene × gene matrices showed any enrichment for gene co-regulation patterns. The co-regulation was calculated based on the permutations (see section **Conversion of *trans*-eQTL results to *trans*-eQTL gene × gene P-value matrices** above). We calculated enrichment for each of the three *trans*-eQTL gene × gene matrices using a contingency table as described above (**row 4 in Supplementary Figure 9**).

Enriched overlap with protein-protein interactions

Next, we calculated enrichment of protein-protein interaction, using information from InBio[106], downloaded from https://www.intomics.com/inbio/map/api/get_data?file=InBio_Map_core_2016_09_12.tar.gz (**row 5 in Supplementary Figure 9**).

Enriched overlap with Hi-C contacts

Last, we investigated if there was any enrichment of physical contact (as measured by Hi-C data) within the *trans*-eQTL gene × gene matrix. We used both inter- and intrachromosomal data derived from LCL samples (GM12878, GEO accession GSE63525)[114]. We looked at a resolution of 10 kb, applied KR normalisation[115] using the supplied information and only took contacts with a mapping quality of ≥ 30 (MAPQGE30). We converted this matrix into a gene × gene matrix by assigning the 10 kb blocks to all 19,942 genes if they overlapped (part of) the gene or the 50 kb window around each gene. The gene-gene matrix was filled with the maximum Hi-C contact value noted for the blocks assigned to this gene-gene combination. For testing the enrichment, we divided the gene-gene Hi-C matrix into two states: ‘no contact’ if the value was 0 ‘contact’ if it was >0 to generate another contingency table (**row 6 in Supplementary Figure 9**).

Supplementary Results

Meta-analyses on local and distal gene expression

eQTLs and eQTS are concordant between platforms

Our consortium contains 37 datasets profiled using different expression profiling platforms, including several Illumina and Affymetrix expression array versions (Affymetrix HuEx 1.0-ST and Affymetrix U219) and RNA-seq, making a direct meta-analysis impossible. We therefore made use of co-regulation patterns between genes to assign the best-matching expression probe from each expression array type to each gene (**Methods**). After applying this method, we meta-analysed the different expression profiling platforms on gene-level. We then performed eQTL and eQTS discovery and replication analyses between each combination of platforms. Because the different platforms had variable sample sizes, which resulted in differences in replication power, replication rates varied from 86.3% (among *cis*-eQTLs in the largest replication dataset) to 13% (among *trans*-eQTLs in the smallest replication dataset) **(Supplementary Figure 1A-C**). However, effects that were replicated (FDR<0.05) showed consistent allelic directions for *cis*-eQTLs (average over all comparisons 93.23%), *trans*-eQTLs (average over all comparisons 99.2%) and eQTS (average over all comparisons 99.4%). This demonstrates that our integration method enabled us to combine different expression profiling platforms and, importantly, that the eQTLs and eQTSs identified by our approach are replicable between different whole blood datasets (**Supplementary Figure 1A-C**). Therefore we continued with the combined meta-analysis.

Multiple testing correction

As our analysis tested nearly 20,000 genes, our study required a strategy to correct for multiple testing. Bonferroni correction is overly stringent for eQTL analysis due to many correlating genes and extensive linkage between genetic variants. Instead, permutation-based approaches[91,116,117] or Benjamini-Hochberg FDR[93,118,119] are often used for multiple testing correction in eQTL studies. Here, we adopted a permutation-based strategy[91,116,120] where each cohort performed the regular analyses and 10 permutations in which the links between gene expression and genotypes were shuffled in each permutation (**Methods**). As with the non-permuted results, we meta-analyzed the results from each permutation and compared the P-value distributions across all tests between the non-permuted and permuted data to determine an FDR estimate for each association (methodology varies slightly between *cis*-eQTL, *trans*-eQTL and eQTS analyses, see details in **Methods**). We have previously shown that these FDR estimates stabilize after only a few permutations, demonstrating that 10 permutations is sufficient[116]. By evaluating the FDR estimates over all tests performed, our approach yields an analysis-wide estimate of FDR (i.e. genome-wide for *cis*-eQTLs), rather than a specific FDR estimate per gene, which would require many more permutations. For all the discovery analyses, we observed that our strategy was more conservative than Benjamini-Hochberg FDR and less stringent than the Bonferroni method (**Supplementary Figure 2**). Because users of our resource may require different levels of stringency, we provide both permutation-based FDRs and Bonferroni-corrected P-values for all the reported effects.

Correction for unknown confounders

In all the analyses, we accounted for unknown technical confounders (such as batch effects) and biological confounders (such as inter-individual differences in cell-type-composition) by correcting the expression data per cohort for up to 25 expression principal components (PCs) that were not associated with genetic variation (**Methods**). This correction adjusted for the majority of cell-type-composition effects in a subset of samples from the BIOS cohort (N up to 3,831, **Supplementary Figure 3**, see details in **"**Cell-type-composition effects of *trans*-eQTLs and eQTS"). Nevertheless, we acknowledge that our dataset may still include residual cell-type-composition effects.

Power analyses

We performed power calculations using the pwr v1.3-0 package available for R version 4.0.0. First, we assumed that all variants and genes were present in all datasets, giving a sample size of 31,684. Next, we determined the power to detect an effect size in our discovery meta-analysis. We varied the effect size between 0 and 0.5 in steps of 0.001, and calculated the statistical power using the pwr.r.test function, using significance thresholds corresponding to an FDR<0.05 in the *cis*-eQTL ($P<2.0\times10^{-5}$), *trans*-eQTL ($P<8.3\times10^{-6}$) and eQTS ($P<3\times10^{-6}$) analyses.

We then determined the minimal effect size (r=0.024, r=0.025, and r=0.028 for *cis*-eQTLs, *trans*-eQTLs and eQTS, respectively), median effect size (r=0.124, r=0.033, and r=0.037) and maximal effect size (r=0.91, r=0.541 and r=0.225) at the FDR<0.05 significance level (**Supplementary Figure 8A**, **E** and **I**). The observed minimal effect sizes were comparable to the effect sizes at a power of 80%, given the sample size of 31,684: 0.029 and 0.03 for *cis*-eQTLs and *trans*-eQTLs, respectively. For eQTS sample size of 28,158 the corresponding effect size was 0.033 (**Supplementary Figure 8B**, **F** and **J**). At the minimal effect sizes, we observed reasonable levels of power: 50.8% for *cis*-eQTLs, 49.9% for *trans*-eQTLs, and 49.9% for eQTS. Median and maximal effect sizes all had a power > 90%.

Next, we determined the minimal sample size required to detect the minimal, median and maximal effect sizes identified in the discovery meta-analysis at the FDR<0.05 significance thresholds. At a power of 80%, the sample sizes required to detect the median effect size observed in our discovery meta-analysis was 1,685 for *cis*-eQTLs, 25,407 for *trans*-eQTLs, and 22,595 for eQTS (**Supplementary Figure 8C**, **G** and **K**). We note that at stricter significance thresholds (e.g. when using the Bonferroni correction), these sample sizes will be higher.

The datasets we included for replication purposes all had a smaller sample size (N=1,480 for CD14+ cells) than our discovery analysis, either because they were from purified cell types, they were from single cell studies, or because they were from tissues that are harder to investigate than whole blood. Considering the relatively small size of the effects observed in our discovery meta-analysis, and the relatively small sample sizes of the replication datasets, we expected that many of the effects identified in our meta-analysis would not replicate significantly. To quantify this, we determined the number of *cis*-eQTLs, *trans*-eQTLs and eQTS effects that we would expect to be able to replicate at a power of 80% (**Supplementary Figure 8D**, **H** and **L**). We used the effect sizes r and FDR<0.05 significance thresholds as detected in the discovery *cis*-eQTL, *trans*-eQTL and eQTS meta-analysis, and calculated the power to detect each of the effects using the replication dataset sample size. For the *cis*-eQTLs, we expected to be able to detect at least 20% of the effects detected in the meta-analysis, and at most 47% at a power of 80% (when not accounting BIOS methylation data which formed the subset of discovery analaysis samples). For the *trans*-eQTLs, expected replication was markedly lower, due to the smaller median sample size: we expected at least 0.08% to be replicated, and at most 1.29%. Finally, for eQTS associations, we expected at least 0%, and at most 3.36% to replicate at a power of 80%.

We note however, that these power calculations are hypothetical, and assume no bias on the effect sizes introduced by biological and technical confounders that are often present in eQTL datasets (either in the discovery meta-analysis or the replication datasets), including possible tissue or cell type effects. Consequently, while this analysis may indicate that a certain replication dataset has statistical power to replicate a given effect, this does not guarantee that this effect will be detectable in that replication dataset. Nevertheless, these results indicate that *trans*-eQTL and eQTS effects are generally smaller than those observed for *cis*-eQTLs, and thus require larger sample sizes for formal replication.

Local genetic effects on blood gene expression

Capture Hi-C overlap for *cis*-eQTLs

To assess whether *cis*-eQTL lead SNPs overlapped with chromosomal contact as measured by Hi-C data we used promoter capture Hi-C data downloaded from ChiCP[121] (https://www.chicp.org/) and investigated whether lead *cis*-eQTL SNPs overlap with Hi-C contacts more than expected by chance (**Methods**).

Both, long-range (>100kb) and short-range (<100kb) *cis*-eQTLs were significantly enriched for Hi-C contacts (>100kb: real overlap 27.8%, flipped overlap 13.6%, test of equal proportions $P=3.3\times10^{-12}$; <100kb: real overlap 17.0%, flipped overlap 12.9%, test of equal proportions $P=9.1\times10^{-16}$). We observed that long-range *cis*-eQTLs were more strongly enriched for Hi-C overlap (2.0-fold for >100kb versus 1.3-fold for <100kb).

One third of trait-associated variants have distal effects

Setup of the *trans*-eQTL analysis

An alternative strategy to gain insight into the molecular functional consequences of disease-associated genetic variants is to ascertain *trans*-eQTL effects. Genome-wide eQTL analyses would have imposed an extensive computational burden on participating cohorts with additional logistic challenges which would arise from the need to share 37 very large summary statistics files (~20,000 genes times ~11M SNPs) with central site. Therefore, we constrained our analyses to a subset of 10,317 variants that have previously been associated with complex phenotypes (**Methods**, **Supplementary Table 3**).

Effects of measured and predicted cell metrics on *trans*-eQTL effects

We aimed to distinguish *trans*-eQTLs caused by intracellular molecular mechanisms from eQTLs induced by blood cell type-composition. To do so, we investigated a subset of up to 1,858 whole blood samples from the BIOS Consortium for which 49 measured and predicted blood cell metrics were available (**Supplementary Methods**, see details in **Supplementary Comment**). We first reasoned that if a *trans*-eQTL is intracellular (i.e. not driven solely by cell-type-composition), the main *trans*-eQTL effect should remain after correcting for cell-type-composition differences. We constructed a linear model incorporating all 49 available cell metrics (**Methods**) and tested whether a residual main effect remained for each *trans*-eQTL. We were able to test 55,311 *trans*-eQTLs in this subset (minor allele frequency (MAF) >0.05 in each BIOS cohort) and found that 4,241 (7.67%) were below the P-value threshold ($P<8.3\times10^{-6}$, threshold determined in discovery meta-analysis) in a linear model without any cell type metrics. Out of these, 2,952 (69.6% of 4,241 effects) *trans*-eQTLs remained below the significance threshold when all 49 cell metrics were included in the model (**Supplementary Figure 19; Supplementary Data 6**). Here we need to acknowledge that cell-type-composition may lead to false positive *trans*-eQTL effects, but we also

note that large-scale cell count measures were not available for many included cohorts, which precluded us from drawing definite conclusions about this issue. We next reasoned that, if a *trans*-eQTL is generic (i.e. it has similar effect sizes within each individual cell type), the main *trans*-eQTL effect would also remain after correcting for cell-type-composition differences and their interactions with the *trans*-eQTL SNP. When we included all the interaction terms between cell-type metric and genetic variant in the model, only 33 (0.06%) out of 4,241 *trans*-eQTLs remained below the P-value threshold ($P<8.3\times10^{-6}$), suggesting that most *trans*-eQTLs have variable effect sizes in different blood cell types (**Supplementary Methods; Supplementary Figure 19**). We also aimed to assign each of the *trans*-eQTLs to the cell type it most likely manifests in by testing the interaction between genotype and each cell metric (**Supplementary Methods**). However, no individual interaction effects were below the FDR threshold (Benjamini-Hochberg FDR>0.05; smallest $P=1.37\times10^{-7}$; **Supplementary Data 7**), likely due to the extensive multiple testing burden and limited power.

Trans-eQTL replication analyses in purified cell types, cell lines, and methylation data

Our replication analyses between different expression platforms suggest that *trans*-eQTLs are replicable between blood datasets (**Supplementary Figure 1B**) but cannot identify cell-type-composition effects. To estimate the fraction of *trans*-eQTLs that constitute intracellular *trans*-eQTLs, we performed replication analyses in bulk RNA-seq datasets derived from specific cell types: lymphoblastoid cell lines (LCL), induced pluripotent cells (iPSCs) and several purified blood cell types (CD4+, CD8+, CD14+, CD15+/CD16+, CD19+, monocytes and platelets). Additionally, we used blood DNA methylation QTL data to support the validity of *trans*-eQTLs. In total, 4,018 (6.7% of the total) *trans*-eQTLs showed replication in at least one cell type (Benjamini-Hochberg FDR<0.05; 93.3% with same allelic direction, on average) or were supported by the methylation data (Benjamini-Hochberg FDR<0.05; meQTL effect direction supporting the discovery eQTL effect, see **Methods**, **Supplementary Figure 5**, **Supplementary Data 2**).

Trans-eQTL replication analyses in post-mortem tissues

We then investigated whether *trans*-eQTLs are shared across tissues from GTEx[117,122]. We repeated our discovery meta-analysis while excluding whole blood samples from GTEx, performed replication analyses in all GTEx tissues, and observed that the replication rate was very low (0.07% of *trans*-eQTLs replicated in any non-blood tissue, 0.09% in blood, Benjamini-Hochberg FDR<0.05). However, the allelic concordance of significant effects was, on average, 66% in non-blood tissues and 100% in blood (**Supplementary Data 3**). Despite these low replication rates, *trans*-eQTLs showed an inflation of replication signal in the majority of tissues (**Supplementary Figure 6A**), most notably in whole blood, esophagus muscularis, liver, heart atrial appendage and non-sun-exposed skin.

Trans-eQTL replication analyses in scRNA-seq data

Ideally, replication of individual *trans*-eQTLs should be performed using single-cell (sc)RNA-seq eQTL datasets, since such datasets are less impacted by the cell-type-composition differences present in bulk eQTL datasets. Currently available scRNA-seq eQTL datasets are still relatively small, but by meta-analysing two different PBMC-based scRNA-seq cohorts using the 10X Chromium platform (OneK1K, N=982 and 1M-scBloodNL, N=157), we were able to perform *trans*-eQTL replication analysis in B-cells, CD4+ T-cells, CD8+ T-cells, classical monocytes, non-classical monocytes, dendritic cells, natural killer (NK) cells and plasma cells from up to 1,139 individuals (up to 3.6% of the discovery sample size, see further information in **Replication Datasets** and **Supplementary Comment**). For each of the 59,786 discovery *trans*-eQTLs, we tested the association within each cell type, but only if the *trans*-eQTL gene was sufficiently expressed (i.e. had a missing sample fraction of at most 20% in the larger OneK1K dataset). We

did this because the expression of only a few thousand genes per cell were quantified in scRNA-seq data.

Since scRNA-seq eQTL data is noisier than bulk RNA-seq data, fewer eQTLs can be identified when using the same number of samples[123]. Moreover, *trans*-eQTLs in eQTLGen were identified using 31,684 samples, while the single-cell replication cohort was limited to 1,139 individuals. Therefore, since the statistical power to formally replicate *trans*-eQTLs was limited, we first studied whether there was any inflation of replication test statistics. For 7 out of the 8 cell types examined, we observed inflation of signal (**Supplementary Table 5**, **Supplementary Figure 6A**; for the least abundant cell type, plasma cells (**Figure 3A**), no inflation of signal was observed) and greater than expected allelic concordance with the discovery analysis (**Figure 3A; Supplementary Table 5;** two-sided binomial test P<0.05). Similarly, by correlating the effect sizes of independent *trans*-eQTLs using the r_b method (**Methods**)[124], we observed that blood *trans*-eQTL effect sizes correlate significantly with replication effects in the scRNA-seq data (**Figure 3A**; **Supplementary Table 5**; two-sided P<0.05) for 4 out of 8 cell types (classical monocytes ($P=3.36\times10^{-8}$, $r_b=0.514$, S.E.=0.093), NK cells ($P=3.24\times10^{-4}$, $r_b=0.185$, S.E.=0.051), CD8+ lymphocytes ($P=3.41\times10^{-3}$, $r_b=0.454$, S.E.=0.155) and B cells ($P=5.98\times10^{-3}$, $r_b=0.049$, S.E.=0.018)). More abundant cell types showed higher *trans*-eQTL effect size correlations with whole blood (**Figure 3A**, Pearson $R^2=0.53$, two-sided P=0.04). When conducting r_b analysis on the bulk expression profiles from purified blood cell types (**Supplementary Figure 7**; average $r_b=0.55$), we observed r_b metrics similar to scRNA-seq data for several cell types, demonstrating that there is concordance between scRNA-seq and bulk expression data from specific cell types.

These correlations and inflations of signal show that some of the *trans*-eQTLs identified in blood are also present in the cell types in our scRNA-seq data, although it remains challenging to prioritize individual effects. Still, we aimed to formally replicate individual *trans*-eQTLs. Depending on the cell type, we could reliably test between 1,917 and 27,582 of the *trans*-eQTLs identified in the discovery analysis (**Figure 3A**). We replicated 35 *trans*-eQTLs at FDR<0.05 (**Supplementary Table 4**), with two effects appearing in more than one cell type. For *trans*-eQTLs which replicated, the allelic concordance between the discovery and the replication analysis was very high (97% concordance), providing additional support for valid replication of these eQTLs.

Lastly, to increase the statistical power to replicate individual *trans*-eQTLs in the noisy scRNA-seq data, we combined the summary statistics from 8 cell types by averaging the Z-scores per *trans*-eQTL over the available cell types. When confining the analysis to the 729 *trans*-eQTLs with an absolute average Z>1.96 (corresponding to a nominal P<0.05, **Supplementary Table 4**), we observed a relatively high concordance of 84% (**Figure 3A**, **Supplementary Table 5**, two-sided binomial test; $P=1.25\times10^{-84}$) suggesting that many of these *trans*-eQTLs represent effects that are independent of cell-type-composition. Among the 729 *trans*-eQTLs, we observed a strong enrichment for genes involved in cytokine-mediated signalling (hypergeometric test from ToppGene[125], $P=3.3\times10^{-12}$, Benjamini-Hochberg FDR<0.05).

The overlap between *cis*- and *trans*-eQTLs

To evaluate which *trans*-eQTL SNPs also have *cis*-eQTL effects, we conducted locus-wide *trans*-eQTL analyses in a subset of samples (N=4,339; EGCUT and BIOS cohorts; **Supplementary Figure 17**; **Supplementary Methods**). For this analysis, we focused on *trans*-eQTLs identified in the discovery meta-analysis. We extracted the *trans*-eQTL SNPs that showed significant effect in this subset of samples ($P<8.3\times10^{-6}$; P-value threshold estimated using discovery *trans*-eQTL meta-analysis) and constructed 12,911 *trans*-eQTL loci (SNP - *trans*-eQTL gene combination; ±1 Mb from tested GWAS SNP) (**Methods, Supplementary Figure 17**). We then performed iterative conditional *trans*-eQTL analyses to identify independent lead *trans*-eQTL SNPs for each locus (**Supplementary Table 21**). For each of these lead *trans*-eQTL SNPs, we then calculated linkage disequilibrium (LD) with lead *cis*-eQTL SNPs identified in the discovery meta-analysis. Out of 12,911 *trans*-eQTL loci, 3,786 (29.3%) were in LD with at least one lead *cis*-eQTL SNP ($R^2>0.8$

between *cis*-eQTL and *trans*-eQTL lead SNPs, 1kG p1v3 EUR, **Supplementary Tables 22-23**). Since the discovery *cis*-eQTL and *trans*-eQTL analyses were performed in the same set of samples, we note that this estimated proportion might be somewhat biased. However, corresponding *cis*-eQTL genes were strongly enriched for having transcription factor (TF) activity ("RNA polymerase II regulatory region sequence-specific DNA binding (GO:0000977)"; one-sided Fisher's exact test $P=9.15\times10^{-6}$, Benjamini-Hochberg FDR=0.043; **Supplementary Figure 20**).

Interaction analyses between *cis*- and *trans*-eQTLs

These LD-based lead-SNP-overlap analyses identify loci where two association signals likely overlap. We next formally tested whether local genes within 100kb of the *trans*-eQTL SNP affect the expression of the *trans*-eQTL gene, limiting the analysis to non-HLA *trans*-eQTLs detected in the discovery meta-analysis. We used a subset of 4,339 samples from the BIOS and EGCUT cohorts and included the local gene in a linear model as a gene-environment (G × E) interaction term. We considered *trans*-eQTLs with a Benjamini-Hochberg FDR<0.05 for an interaction term to be driven by the expression of a *cis*-acting gene. We observed interaction effects for 615 out of 201,106 SNP–*cis*–*trans*–gene combinations tested (**Supplementary Table 24**), reflecting 585 *trans*-eQTLs. For instance, for rs7045087 (associated to red blood cell counts[126]), we observed that the expression of the interferon gene *DDX58* (mapping 38bp downstream from rs7045087) interacted with *trans*-eQTL effects on *HERC5, OAS1, OAS3, MX1, IFIT1, IFIT2, IFIT5, IFI44, IFI44L, RSAD2* and *SAMD9* (**Supplementary Figure 18**), most of which are involved in interferon signaling. These results indicate that *trans*-eQTL effects can be affected by the expression of local genes, but comprehensive characterization of such interaction effects requires larger sample sizes.

Biological mechanisms leading to *trans*-eQTLs

We studied the biological nature of the *trans*-eQTLs we identified. To do this, we assessed whether *trans*-eQTLs were enriched for TF - downstream target pairs, co-regulated genes, protein-protein interactions (PPI) and inter-chromosomal contacts. To perform these analyses, we converted the *trans*-eQTL SNP-gene matrix into three *trans*-eQTL gene × gene matrices (**Supplementary Methods**).

We observed that there is significant enrichment of TF - downstream target pairs[113] among our three *trans*-eQTL matrices with odds ratios (ORs) up to 1.40 for the combined *trans*-eQTL matrix (**Supplementary Figure 9, row 1**).

We observed stronger and highly significant enrichments when increasing the list of TF-downstream target gene pairs with genes co-regulated with known TFs (OR=1.38, $P=5.8\times10^{-72}$, **Supplementary Figure 9, row 3**), or with known TF-target genes (OR=3.57, $P<1\times10^{-308}$, **Supplementary Figure 9, row 2**), or with both (OR=4.37, $P<1\times10^{-308}$, **Supplementary Figure 9, row 4**).

When using co-regulation to predict relationships between genes based on expression patterns we observed an overlap with 797 significant *trans*-eQTL gene pairs, reflecting a 22.3-fold enrichment ($P<10^{-308,}$ **Supplementary Figure 9, row 5**).

Significant *trans*-eQTL gene-pairs are also enriched for interactions among proteins: 584 protein-protein interactions (from a total of 799,176 protein-protein interactions that were derived from InWeb[111]) overlapped with our *trans*-eQTLs ($P<10^{-17}$, **Supplementary Figure 9, row 6**). Some of these pairs consist of genes that encode subunits of the same protein complex (e.g. *POLR3H* and *POLR1C*).

Lastly, there was a highly significant enrichment ($P=2.4\times10^{-153}$) for inter- and intra-chromosomal contacts when using Hi-C data[114]: the OR was 1.47 for the combined *trans*-eQTL matrix.

These enrichments show that *cis*-eQTLs are informative for understanding *trans*-eQTLs, but that including blood-derived *cis*-eQTLs do not capture all locally informative effects. In most of these analyses, we found stronger enrichments in the combined *trans*-eQTL matrix (third column in **Supplementary Figure 9**), as compared to the matrix created by only including local SNPs (Pascal method, first column in **Supplementary Figure 9**) or by only including *cis*-eQTL SNPs (second column in **Supplementary Figure 9)**. Putative mechanisms for each of the 59,786 *trans*-eQTLs are listed in the **Supplementary Data 4**.

Per-phenotype enrichment analyses for *trans*-eQTLs

Next, for each GWAS phenotype, we interrogated whether *trans*-eQTL genes were enriched for Gene Ontology (GO) terms. In total, we observed 347 enriched GO terms for 208 out of 345 (60%) traits (one-sided Fisher's exact test, Benjamini-Hochberg FDR<0.05; **Supplementary Table 12**). We observed that several of the enriched GO terms were relevant for the tested trait. For example, *trans*-eQTL SNPs associated with celiac disease and inflammatory bowel disease showed the strongest enrichments for GO terms associated with response to cytokine stimulus (e.g. celiac disease: "cellular response to cytokine stimulus", FDR=1.06×10^{-5}), platelet count was enriched for "platelet degranulation" (FDR=2.6×10^{-10}), and heart rhythm traits were most enriched for cholesterol-related terms (e.g P-wave duration was enriched for "regulation of cholesterol biosynthetic process", FDR=4.6×10^{-14}).

Examples of *trans*-eQTLs

In the following examples we highlight *trans*-eQTLs where the eQTL SNP influences the gene expression level through various mechanisms.

Some *trans*-eQTLs can influence genes strongly expressed in tissues other than blood. For example, rs17087335 (dbSNP 137, associated with coronary artery disease[127]) affects the expression of 88 genes in *trans* (FDR<0.05, Bonferroni corrected P<0.05 for 39 genes; **Figure 4, Supplementary Table 25**) that are highly expressed in brain (one-sided Fisher's exact test, ARCHS4 database, Benjamini-Hochberg FDR=6.43×10^{-14}; **Figure 4**). Eighty-five out of the 88 (96.6%) *trans*-eQTL genes were upregulated by the minor allele of rs17087335 and strongly enriched for the targets of REST (RE-1 silencing transcription factor; one-sided Fisher's exact test for ENCODE[100,128] project REST ChIP-seq, Benjamini-Hochberg FDR=8.84×10^{-38}, **Figure 4**). While the minor allele of rs17087335 was associated with lower expression of *REST*, it was not in LD (R^2<0.2, 1kG p1v3 EUR) with the lead *cis*-eQTL SNP (rs13353552; dbSNP 137). A SNP in high LD with rs17087335, rs3796529 (R^2=0.91, 1kG p1v3 EUR; dbSNP 137), is a missense variant for REST, suggesting that these *trans*-eQTLs could also arise from a post-transcriptional mechanism of action. Because REST is a TF that downregulates the expression of neuronal genes in non-neuronal tissues[129,130], we speculate that the observed *trans*-eQTLs reflect the impact of genetic variation on the effectiveness of downregulation, although experimental follow-up is required to confirm this hypothesis. Nevertheless, this example illustrates that blood *trans*-eQTL effects can help to prioritize the putatively causal *cis*-eQTL gene among multiple genes in a locus (here *REST*).

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[131] (dbSNP 137) is in high LD (R^2>0.8, 1kG p1v3 EUR) with the lead *cis*-eQTL effect for a gene encoding transcription factor ZNF131. *Cis*-eQTL and *trans*-eQTL lead SNPs for this locus were in high LD (R^2>0.8, 1kG p1v3 EUR) for 25 out of the 75 distal downstream genes (**Supplementary Figure 11A**). In a recent short hairpin RNA knockdown experiment of *ZNF131*[132], three separate cell isolates showed downregulation of four genes that we identified as *trans*-eQTL genes: *HAUS5, TMEM237, MIF4GD* and *AASDH* (**Supplementary Figure 11A**). ZNF131 has been hypothesized to inhibit estrogen signaling[133], which may explain how the SNP in this locus contributes to altering the age of menarche.

***Trans*-eQTLs extend insight for loci with multiple *cis*-eQTL effects.** In the *FADS1*/*FADS2* locus, rs174574 is associated with lipid levels[5] and affects 17 genes in *trans* (**Supplementary Figure 11B**). The strongest *cis*-eQTLs modulate the expression of *FADS1*, *FADS2* and *TMEM258*, with the latter being in high LD with GWAS SNP (R^2>0.8, 1kG p1v3 EUR). From those genes, *FADS1* and *FADS2* have been implicated[5] since these encode fatty acid desaturases, and consistent with their biological function, *trans*-eQTL genes from this locus are highly enriched for triglyceride metabolism ($P<4.1\times10^{-9}$, GeneNetwork[134] REACTOME pathway enrichment). Since this locus has extensive LD, variant and gene prioritization is difficult: conditional analyses in the subset of 4,339 BIOS and EGCUT samples 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; dbSNP 137). 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 (dbSNP 137) is associated with multiple immune-related traits (type 1 diabetes (T1D), inflammatory bowel disease (IBD), systemic lupus erythematosus (SLE) and psoriasis[41,135-137]). For this SNP we identified 17 *trans*-eQTL effects, but no detectable gene-level *cis*-eQTLs in blood (**Supplementary Figure 11C**). 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 which might lead to the observed *trans*-eQTLs. Connection between MDA5 and T1D has been previously described[138]. 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 26**).

***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[24] (dbSNP 137) has 14 *cis*-eQTL effects, most notably on *IKZF3*, *GSDMB*, and *ORMDL3* (**Supplementary Figure 11D**), making it difficult to identify most likely causal gene. 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[96] (**Supplementary Figure 11D**). *IKZF3* is part of the Ikaros transcription factor family that regulates B-cell proliferation[96,139], 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.

***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: on chr 4q12 (**Supplementary Figure 11E**), height-associated SNP rs13113518[29] (dbSNP 137) is 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[140]. Out of seven *trans*-eQTL genes, three were known circadian rhythm genes (*TEF, NR1D1* and *NR1D2*) and showed increased expression for the trait-increasing allele, suggesting a possible mechanism for the observed *trans*-eQTLs through binding of CLOCK:BMAL1. Two out of four remaining *trans*-eQTL genes (*C10ORF116* and *RP1196C23.8*) showed increased expression and two (*ATP1A1* and *KLRB1*) showed decreased expression. *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[141]**.** *NR1D1* and *NR1D2* encode the transcriptional repressors Rev-ErbA alpha and beta, respectively, and form a negative feedback loop to suppress *BMAL1* expression[142]. *NR1D1* and *NR1D2* have been reported to be associated with osteoblast and osteoclast functions[143], revealing a possible link between circadian clock genes and height.

eQTSs identify potential driver genes for polygenic traits

Effect of GWAS P-value threshold on eQTS detection

When calculating PGSs, the P-value threshold for including the SNPs that corresponds to most explained variation is likely to be trait-dependent. We therefore calculated PGSs using clumped GWAS lead SNPs at five significance levels (P<0.01; 1×10^{-3}; 1×10^{-4}; 1×10^{-5}; 5×10^{-8}). While we could detect the majority of eQTSs (70.5%) at the most conservative threshold (P<5×10^{-8}), the total number of results was higher than for each P-value threshold separately (**Supplementary Table 27**), suggesting that our analysis captured different genetic architectures. Unsurprisingly, we identified more eQTSs for GWAS with larger sample sizes (Spearman r=0.42–0.59 at different P-value cut-offs). Traits with few eQTS associations typically also had lower average (Spearman r=0.42–0.72) and maximum eQTS effect sizes (Spearman r=0.69–0.85; **Supplementary Table 28**).

eQTS cross-platform replication analyses

As in the previous analyses, the cross-platform replication rates showed high allelic concordance between blood datasets (average concordance rate was 99.2% for effects reaching FDR<0.05 in replication dataset, **Supplementary Figure 1C**), although the replication rates were quite low in the platforms with fewer samples (21.35-26.4% of tested effects reached FDR<0.05 in 1,549 FHS samples, **Supplementary Figure 1C**).

Effects of measured and predicted cell metrics on eQTS effects

Similar to our analysis of *trans*-eQTLs, we investigated whether eQTS could be driven by interindividual differences in cell-type-composition. We fitted linear models with and without cell-type metrics as covariates in a subset of 1,858 samples (**Supplementary Methods**). Out of 18,210 eQTSs, 2,313 (12.7%) were below the P-value threshold in the original model (P<3.02×10^{-6}, threshold determined by discovery meta-analysis). When all 49 cell metrics were included, 618 (3.39%) out of 2,313 eQTSs remained below the P-value threshold (**Supplementary Table 29**, **Supplementary Figure 19**). Twenty-one (3.4%, affecting 7 genes) replicated in at least one of our replication datasets. However, the majority of replicating effects originated from PGSs of erythrocyte- and platelet-related GWAS traits, while also affecting several blood-related genes such as *HBG1* and *HBG2*. This suggests that some strong cell-type-composition effects might still be detectable after correcting the data for all the main effects. When including all interaction terms between cell-type metric and PGS, only two eQTSs (0.01%) remained below the P-value threshold (P<3.02×10^{-6}), demonstrating the cell-type-specific nature of eQTSs. In line with the *trans*-eQTL effects, none of the eQTS effects could be reliably assigned to any of the cell-type metrics when testing individual PRS-cell metric interaction effects (**Supplementary Methods**, Benjamini-Hochberg FDR>0.05; smallest P=1.31×10^{-6}; **Supplementary Table 30**).

eQTS replication analyses in cell lines and post-mortem tissues

We next ascertained to what extent eQTS associations can be replicated in independent datasets by studying 1,460 LCL samples, 762 iPSC samples and all GTEx tissues[117]. We were able to replicate 10 eQTSs in the LCL dataset, and 9 out of 10 (FDR<0.05) had the same effect direction as in the discovery dataset (**Supplementary Figure 12A, Supplementary Data 5**). Seventy-eight eQTSs replicated in the iPSCs dataset (FDR<0.05), with 71 (91%) showing the same direction of effect (**Supplementary Figure 12B, Supplementary Data 5**). Since polygenic risk scores can differ substantially between populations, we performed GTEx replication analyses while confining ourselves to Europeans and identified 19 replicating eQTSs with FDR<0.05 and same direction of effect (eQTS discovery performed without GTEx; 66 replicated when also including non-

European samples, **Supplementary Data 6-7**). We observed the inflation of replication signal in some tissues, primarily in blood (**Supplementary Figure 6B**).

Because only a few eQTS associations were replicated, there was no strong replication signal in non-blood tissues, and the majority of identified eQTS associations were observed for blood-related traits (**Extended Data Figure 3, Supplementary Data 5**), we speculate that these effects are highly tissue- or cell-type-specific. However, as suggested by the power analyses, the limited replication in other tissues could also be a result of the small effect size of eQTS effects (median r=0.037; **Supplementary Figure 8I**) causing a lack of statistical power in the replication datasets due to their small sample size, or because of variability in PGS estimates caused by differences in sample characteristics (e.g. age, sex, socio-economic status, etc) of the included datasets[144].

Features of eQTS genes

We took 2,568 significant eQTS genes (FDR<0.05) and conducted over-representation analysis (one-sided Fisher's exact test) for Gene Ontology (GO) categories (**Supplementary Table 14**). While we observed 51 (1.08%) significant GO terms (Benjamini-Hochberg FDR<0.05), none implicated transcriptional regulation (e.g. "RNA polymerase II core promoter sequence-specific DNA binding", FDR=0.52). Some of the most significant terms were associated with cellular secretion (e.g. "secretory granule lumen", FDR=2.2×10^{-8}), blood cell traits ("platelet degranulation", FDR=3.4×10^{-6}, "regulation of B cell proliferation", FDR=5.8×10^{-4}) and intercellular signalling ("pattern recognition receptor signaling pathway"; FDR=5.2×10^{-4}).

Similarly, we tested for enrichment of 1,672 known TFs from the FANTOM5 database and did not observe any enrichment (one-sided Fisher's exact test; unadjusted P=0.99, OR=0.56): out of 2,568 eQTS genes, 140 (5.45%) overlapped with known TFs.

To test whether eQTS genes have relatively more interaction partners in protein–protein networks, we used the protein-protein interaction (PPI) data from InWeb[111]. After intersecting the dataset with genes tested in eQTLGen (19,942), 13,355 remained, in which 2,132 were eQTS genes (FDR<0.05 in the discovery meta-analysis). We compared the numbers of PPI interaction partners between eQTS genes and non-eQTS genes using the Wilcoxon-Mann-Whitney test and observed that non-eQTS genes had significantly more interaction partners (two-sided Wilcoxon-Mann-Whitney test, P=1.98×10^{-5}; average 69.38, median 25.00) than eQTS genes (average 58.41, median 20).

These analyses suggest that eQTS genes are not enriched by TF targets or protein–protein hubs and that such genes may have variety of mechanisms for regulation.

Per-phenotype enrichment analyses for eQTS genes

When stratifying eQTS effects by GWAS phenotype, we identified 90 phenotypes showing enrichment with any GO term (one-sided Fisher's exact test, Benjamini-Hochberg FDR<0.05; **Supplementary Table 31**), and these often reflected the known biology. For instance, eQTS genes for platelet count showed the strongest enrichment for the process "platelet degranulation" (FDR=6×10^{-17}), monocyte count for "neutrophil degranulation" (FDR=4.7×10^{-16}) and eQTS genes for total lipids in large HDL for "cholesterol metabolic process" (FDR=1.6×10^{-6}).

Examples of eQTSs

In the next sections we highlight several eQTSs demonstrating the insights that can be gained from eQTS analysis.

eQTS analysis identified genes relevant for non-blood traits. As an example, the association of *GPR15* (P=3.7×10^{-8}, FDR<0.05; **Supplementary Figure 13A**) with the trait 'ever versus never

smoking'[81]. *GPR15* is a biomarker for smoking[145] that is overexpressed and hypomethylated in smokers[146]. We observe strong *GPR15* expression in lymphocytes (**Supplementary Figure 13A**), suggesting that the association with smoking could originate from a change in the proportion of T cells in blood[147]. As *GPR15* is involved in T cell homing and has been linked to colitis and inflammatory phenotypes, it is hypothesized to be involved in the systemic inflammation induced by tobacco smoking[147].

The PGS for another non-blood trait, educational attainment[80], correlated significantly with the expression of 21 genes (FDR<0.05; **Supplementary Figure 13B**, **Supplementary Table 32**). Several of the strongly associated genes are known to be involved in neuronal processes (**Supplementary Figure 13B**, **Supplementary Table 32**) and show expression in neuronal tissues (GTEx v7, **Supplementary Figure 13C**)[148]. *STX1B* (strongest eQTS $P=1.3\times10^{-20}$) is specifically expressed in brain (**Supplementary Figure 13C**), and its encoded protein, syntaxin 1B, participates in the exocytosis of synaptic vesicles and synaptic transmission[149]. Another gene highly expressed in the 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[150]. 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 with synaptic plasticity[151]. *NRG1* was also positively associated with the PGS for monocyte levels[86] (strongest eQTS $P=1.5\times10^{-7}$), several LDL cholesterol traits (e.g. medium LDL particles[30]; strongest eQTS $P=6.2\times10^{-8}$), coronary artery disease[127] (strongest eQTS $P=1.5\times10^{-6}$) and body mass index in females[27] (strongest eQTS $P=9.2\times10^{-12}$).

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 33**). For example, the interferon gene *STAT1* was significantly associated with T1D, celiac disease (CeD), inflammatory bowel disease (IBD) and primary biliary cirrhosis (PBC) PGSs. However, some of these genes are also marker genes for specific blood-cell types, such as *CD79A*, which showed a significant correlation with type I diabetes (T1D) and PBC. To test whether disease-specific eQTS gene signatures are reflected by blood cell proportions, we investigated single-cell RNA-seq data[96] (**Methods**; **Supplementary Figure 13D**). For ulcerative colitis (a subtype of IBD), we observed a significant depletion of expression in megakaryocytes. SLE eQTS genes were enriched for antigen presentation (GeneNetwork[134] $P=1.3\times10^{-5}$) and interferon signaling (GeneNetwork $P=1.4\times10^{-4}$), consistent with the well-described interferon signature in SLE patients[153,153]. Moreover, the SLE genes were significantly enriched for expression in myeloid dendritic cells, whose maturation depends on interferon signaling[154]. 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.

Supplementary Equations

Here we theoretically evaluate the potential correlation between polygenic score PGS and expression of a gene j, using the model and notation from the Liu/Li/Pritchard 2019 "LLP19" Cell paper[155]. The polygenic score for individual i will be defined as the $E(Y_i - \bar{Y}|G_i)$ where $Y_i$is the phenotype of individual i and $\bar{Y}$ is the mean phenotype, and G_i is the genotype of individual i. Under the LLP19 model, the expected phenotype depends on the expression levels of core genes as follows:

$$PGS_i = E(Y_i - \bar{Y}|G_i) = \sum_{j=1}^{M} \gamma_j \left(x_{i,j} - \bar{x}_j\right) \quad (1)$$

where M is the number of core genes, $x_{i,j}$ is the expression level of gene j in individual i and $\bar{x}_j$ is the corresponding mean for gene j, and γ_j measures the effect of expression of gene j on the phenotype. Note that here we consider an idealized score assuming perfect information about SNP effects on expression; the computed polygenic score is presumably proportional to this with additional error. Then we want to compute

$$Cor\left[x_{i,j}, PGS_i\right] = \frac{Cov\left[x_{i,j} PGS_i\right]}{SD\left[x_{i,j}\right] SD\left[PGS_i\right]} \quad (2)$$

where variances and covariances are computed across individuals (i), with respect to a specified gene j.

The numerator is

$$Cov\left[x_{i,j}, PGS_i\right] = E\left[\sum_{s=1}^{S} \beta_{s,j}\left(g_{i,s} - 2p_s\right)\right] \times \left[\sum_{k=1}^{M}\sum_{s=1}^{S} \gamma_k \beta_{s,k}\left(g_{i,s} - 2p_s\right)\right] \quad (3)$$

where $\beta_{s,j}$ is the effect of SNP s and gene j; $g_{i,s} \in 0, 1, 2$ is the genotype of individual i and SNP s and $2p_s$ is the mean genotype at SNP s; and where s indexes across all S SNPs that are *cis*- or *trans*-eQTLs for *any* core gene. We set $\beta_{s,j}$= 0 for SNP-gene combinations that are not eQTLs. (Note that this is a minor change in notation from LLP19 where we used different SNP indexing for each gene, and listed *cis*- and *trans*-eQTLs separately.) Then if j is a core gene, this can be rewritten as

$$Cov\left[x_{i,j}, PGS_i\right] = \sum_{k=1}^{M} \gamma_k C(j,k) \qquad [j\ not\ in\ core\ set] \quad (4)$$

$$Cov\left[x_{i,j}, PGS_i\right] = \gamma_j \sigma_j^2 + \sum_{k=1;k\neq j}^{M} \gamma_k C(j,k) \qquad [j\ in\ core\ set] \quad (5)$$

We have the denominator from LLP19 as follows:

$$Var\left[x_{i,j}\right] = \sigma_j^2 + \sigma_{j,E}^2 \quad (6)$$

where σ_j^2 and $\sigma_{j,E}^2$ are the genetic and environmental variances in expression of gene j, and

$$Var[PGS_i] = \sum_{k=1}^{M} \gamma_k^2 \sigma_k^2 + \sum_{k=1}^{M} \sum_{l=1,\neq k}^{M} \gamma_k \gamma_l C(k,l) \quad (7)$$

Putting this together, in the case where j is in the cores set we have

$$Cor\left[x_{i,j}, PGS_i\right] = \frac{\gamma_j \sigma_j^2 + \sum_{k\neq j}^{M} \gamma_k C(j,k)}{\left[\sigma_j^2 + \sigma_{j,E}^2\right]^{\frac{1}{2}} \left[\sum_{k=1}^{M} \gamma_k^2 \sigma_k^2 + \sum_{k=1}^{M} \sum_{l=1,\neq k}^{M} \gamma_k \gamma_l C(k,l)\right]^{\frac{1}{2}}} \quad (8)$$

The denominator is a bit of a mess, but we can get a sense of this for special cases:

Suppose that core genes are uncorrelated ($C = 0$), that the γs and σs are all equal, and that $\sigma_{j,E}^2 = 0$. Then the PGS:expression correlation is $1/\sqrt{M}$. More realistically, the nongenetic variance $\sigma_{j,E}^2$may be considerable. In this case the PGS:expression correlation is $\sigma_j^2 / \sqrt{\left(\sigma_j^2 + \sigma_{j,E}^2\right)M}$ which can be rather smaller.

If the correlations are large in the sense that $\sum C \gg \sigma^2$ then (again with the γs and σs all equal, and $\sigma_{j,E}^2 = 0$) we have that the correlation is $\sqrt{C} / \sqrt{\left(\sigma_j^2 + \sigma_{j,E}^2\right)}$ which can be nontrivial: for example if the average genetic covariances between j and all other core genes are 1/100th as large as the total variance of gene j, then the PGS:expression correlation could be as large as 0.1. It is important to note that noncore genes that covary genetically with core genes would also be correlated with PGS under this model.

Note that various factors likely reduce the measured correlation substantially below these idealized predictions: inaccuracies in the PGS estimate; the measured tissue is not an ideal proxy for the tissues/cell types/cell conditions that are most relevant for disease; and as noted above, nongenetic variance in expression of the target gene.

In summary, under the LLP19 model, we expect correlations between the PGSs and the expression of core genes (and co-regulated peripheral genes). In idealized settings, these correlations can be fairly large, but various practical factors will reduce the observed signals below the perfect-data expectations.

Supplementary Comment

Cell type composition effects

Blood tissue is a mixture of cells consisting of a variety of blood cell subpopulations. The amount and composition of individual cell types in blood is affected by genetic[126]. In blood eQTL and eQTS analyses, we have correlated genetic variation and polygenic risk with gene expression measured from the bulk RNA. If the expression of a gene or gene set is considerably higher or lower in a particular cell type as compared to the average expression in blood, it is possible that the *trans*-eQTL or eQTS effect we observe, actually reflects the correlation between genetics and the amount or proportion of specific blood cell type (ccQTL, cell type composition QTL). This is different from a real distal eQTL effect where genetics affect the expression of a distal gene through an intracellular molecular pathway (intracellular eQTL).

In principle, cell type composition effects could also influence *cis*-eQTL analysis, but that would require that a blood-cell-specific gene to map within 1 Mb from the blood-trait-related SNP. For *trans*-eQTLs and eQTSs, on the other hand, each blood-trait-related SNP or eQTS could lead to many ccQTL effects from all over the genome. As such, this mechanism could potentially account for a significant portion of reported *trans*-eQTL and eQTS effects. In fact, these cell type composition effects could potentially be present in any study conducting *trans*-eQTL or eQTS analyses in heterogeneous tissues.

Here, we outline various strategies we used to account for cell type composition effects and to prioritise the effects which are likely not caused by cell type composition. In the discovery *cis*-eQTL, *trans*-eQTL and eQTS analyses, we corrected for non-genetic variation caused by unknown biological and technical factors in each cohort. For that, we calculated first 20 (or 25 for the large BIOS Consortium dataset) gene expression-based principal components (PCs) and, prior any analyses, regressed out from each expression matrix those PCs which were not under genetic regulation. This strategy is in line with previous studies in blood[91,116], and is comparable to using PEER factors[156] for describing and correcting for structure in expression data. The advantage of such methods is that we do not need comprehensive information about technical or biological covariates to account for them, and that is crucial when working with 37 cohorts that each have measured different covariates. We acknowledge that, whereas this strategy should take care of the majority of cell type composition effects, it is not perfect. Our choice of not removing PCs under genetic regulation is meant to limit the removal of real eQTL effects, and differs in that aspect from the choice of removing a fixed number of PCs or PEER factors as has been done in GTEx studies[117]. It is possible that a PC correlates strongly with a specific cell type (e.g. platelet levels). If there is a SNP that also strongly affects platelet levels[126], we would not remove this particular PC. As a result, we could observe *trans*-eQTLs on genes which are more highly or lowly expressed in platelets compared to other cell types in blood.

Here, we used a subset of the eQTLGen consortium data as well as newly generated single cell RNA-seq data to further investigate the effect of blood cell type composition on the *trans*-eQTL and eQTS analyses.

Evaluation of PC correction of gene expression

To determine the consequences of interindividual cell count differences on *trans*-eQTLs, we focused on the whole-blood RNA-seq BIOS dataset, consisting of 3,831 individuals, for which the expression data was corrected for the first 25 PCs.

We first determined whether the PCs correlates with cell counts, and observed that the first 25 PCs are indeed strongly correlated to cell counts (**Supplementary Figure 3**; maximal Spearman R^2= 0.521; minimal two-sided uncorrected $P<1.1\times10^{-308}$). For example, in this dataset, PC2 correlated (FDR<0.05) with 48 out of 57 available cell-type metrics. We also studied PCs 26 to 50 in order to determine whether we actually should have corrected for these components as well. PCs 26-50 showed significant correlation with some cell-type metrics (most significant at $P=9.34\times10^{-39}$ was between PC39 and CD14+CD16+ intermediate monocytes). However, although these correlations were significant, the proportion of explained variance was rather small (maximum Spearman R^2=0.043; 10 PCs showed nominally significant correlation (uncorrected P<0.05) with any cell metric). Considering the weak correlation between cell metrics and PCs 26-50 (which we had not used to correct the expression data), we reasoned that the 25 PCs that we had used for correction were selected appropriately for this dataset. Consequently, for the other datasets in the meta-analysis, most of which have smaller sample size than our in-house dataset, we expect that by adjusting for the first 20 PCs, most of the variation in gene expression levels caused by differences in inter-individual cell type abundances have been adjusted appropriately.

We do note that we did not adjust for PCs which were associated with genetic variants. As a result, 7 out of the first 25 components were not included in the regression in the BIOS dataset. Out of these PCs, we observed the strongest correlation between PC18 and eosinophil percentages (**Supplementary Figure 3;** Spearman R^2=0.23; $P=2.7\times10^{-106}$).

Hence, while the majority of cell type composition effects were likely removed from our meta-analysis by PC correction, we cannot exclude the possibility that some cell-type composition effects remain.

Finding cell type composition effects using blood RNA-seq data

To investigate this further, for each discovery *trans*-eQTL and eQTS, we constructed linear models by using (1) all measured and predicted cell type metrics as covariates or (2) all cell metrics and their interaction terms with genetic variant (or PGS) as covariates.

For those analyses, we used a subset of samples for which 49 cell metrics (out of 56) were available (n=1,858). We had to remove some of the abovementioned cell metrics (LUC_Perc, RBC, RDW, MCH, MPV, MCHC, MCV) and 1,973 samples (out of 3,831) from these analyses due to missingness. We note that two of those cell metrics correlated with PCs which were removed from expression data (e.g. RBC and PC3, Spearman R^2=0.177; MPV and PC1, Spearman R^2=0.32), suggesting that those two cell type effects were adjusted for during PC adjustment stage. As BIOS consists of multiple cohorts, we ran all the analyses per cohort separately, and then conducted a meta-analysis using a weighted Z-score method. Due to limited power to detect distal effects in only the BIOS cohort, we were able to identify 4,241 significant *trans*-eQTLs (7.67% of discovery results) when using only PC-corrected expression and not applying any additional corrections.

We next evaluated whether the addition of 49 cell metrics rendered the remaining SNP or PGS main effect non-significant (significance threshold $P<8.3\times10^{-6}$, determined by discovery meta-analysis). We corrected the gene expression for all 49 cell-type metrics (on top of PC correction) and performed the meta-analysis again. We observed that the majority of the *trans*-eQTLs (2,952; 5.34% of discovery) remained significant (**Supplementary Figure 19**, **Supplementary Data 6**). For the other cohorts we did not have extensive cell-count measures available, and therefore were unable to perform this correction uniformly. However, these results suggest that the gene expression correction procedure (i.e. correcting for 20 expression PCs) that we employed has been quite effective to control for cell-type-composition differences.

This analysis also permitted us to determine to what extent *trans*-eQTLs are cell-type specific. We reasoned that if a *trans*-eQTL is generic (i.e. having similar effect sizes within each individual

cell-type), the main *trans*-eQTL effect should remain when correcting the gene expression of a *trans*-gene for cell-type composition differences as well as it's interaction with the *trans*-SNP (i.e. analogous to[91]). After applying this procedure, we observed that for only 33 *trans*-eQTLs a significant main effect remained ($P<8.3\times10^{-6}$), which suggests that *trans*-eQTLs are highly cell-type specific. We also aimed to assign each of the *trans*-eQTLs to the cell type it most likely manifests in by testing the interaction between genotype and each cell metric. However, likely due to the extensive multiple testing burden and limited power, no individual interaction effects gained significance (Benjamini-Hochberg FDR>0.05; smallest $P=1.37\times10^{-7}$; **Supplementary Data 7**).

For eQTS effects, we observed 2,313 (12.7% of discovery results, **Supplementary Table 29**) significant effects (significance threshold $P<3.02\times10^{-6}$, determined by discovery analysis) in BIOS before cell type composition correction, while 618 (3.4% of discovery) remained significant when adjusting for 49 cell type metrics. When expanding the model to include interaction terms between cell metric and PRS, two eQTS remained significant, demonstrating the cell-type-specific nature of eQTSs. In line with the *trans*-eQTL effects, none of the eQTS effects could be assigned to any of the cell type metrics when testing individual genotype-cell metric interaction effect (**Methods**, Benjamini-Hochberg FDR>0.05; smallest $P=1.31\times10^{-6}$; **Supplementary Table 30**).

Prioritizing *trans*-eQTLs based on bulk transcriptome replications

To further estimate what fraction of identified *trans*-eQTLs constitute intracellular *trans*-eQTLs, we performed thorough replication analyses in bulk gene expression datasets derived from blood and other tissues. As blood-based datasets, we investigated lymphoblastoid cell lines (LCL), induced pluripotent cells (iPSCs), several purified blood cell types (CD4+, CD8+, CD14+, CD15+/CD16+, CD19+, monocytes and platelets), and blood DNA methylation QTL data. In total, 4,018 (6.7% of the total) of *trans*-eQTLs showed replication in at least one cell type (FDR<0.05; same allelic direction) or were supported by the methylation data (FDR<0.05; meQTL effect direction supported the discovery eQTL effect); **Supplementary Figure 5**, **Supplementary Data 2**).

We then investigated whether *trans*-eQTLs are shared across tissues from GTEx. We repeated our discovery meta-analysis while excluding GTEx, and observed that the replication rate was very low (0.07% of *trans*-eQTLs replicated in any non-blood tissue, 0.09% in blood, FDR<0.05, same allelic direction; **Supplementary Data 3**), possibly due to low sample size. However, the allelic concordance of significant effects was, on average, 66% in non-blood tissues and 100% in blood (**Supplementary Data 3**). Despite these low replication rates, *trans*-eQTLs showed an inflation of replication signal in the majority of tissues (**Supplementary Figure 6A**), most notably in whole blood, esophagus muscularis, liver, heart atrial appendage and non-sun-exposed skin.

Single cell RNA-sequencing replication

While the aforementioned results were based on small replication sets, the replication in purified cells and inflation in non-blood tissues indicates that a considerable fraction of the *trans*-eQTL effects are not due to cell-type composition effects. However, as we outlined in the section '**Evaluation of PC correction of gene expression**', it is conceivable that the gene expression data has not been corrected for cell-type composition differences entirely.

The best strategy to determine the cell type specificity of *trans*-eQTLs is to use single-cell eQTL data instead of bulk RNA-seq eQTL data. However, single-cell RNA-seq data on a substantial number of individuals, necessary in order to have sufficient statistical power, did not yet exist when we submitted our original manuscript. To resolve this, we have spent substantial efforts to generate single-cell RNA-seq data on PBMCs from 1,139 unrelated individuals in two cohorts:

OneK1K (N=982) and 1M-scBloodNL (N=157) using the 10X Chromium platform. Details on quality control, normalisation and cell type classification for these datasets are provided in the **Supplementary Methods** and follow the procedures that we employed before in an earlier single-cell eQTL pilot study using 45 individuals[91]. Once we had generated data on the two cohorts, we performed *trans*-eQTL replication in B-cells, CD4+ T-cells, CD8+ T-cells, classical monocytes, non-classical monocytes, dendritic cells, natural killer cells and plasma cells.

For each of the 59,786 discovery *trans*-eQTLs, we tested the association within each of these cell types, but only if the *trans*-eQTL gene was sufficiently expressed (i.e. had a missing sample fraction that was at most 20% in the large OneK1K dataset). This was due to the fact that in single-cell RNA-seq data, the expression of only a few thousand genes per cell were quantified.

Since single-cell RNA-seq eQTL data is noisier than bulk RNA-seq, fewer *cis*-eQTLs can be identified when using the same number of samples as we previously observed when studying 45 individuals[14]. Moreover, since *trans*-eQTLs in eQTLGen had been identified using a sample size of 31,684 and our single-cell replication cohort was limited to 1,139 individuals, statistical power was not sufficient to replicate many *trans*-eQTLs. We therefore first studied whether inflation of test statistics was observed. This indeed was the case for 7 out of the 8 cell-types studied (**Supplementary Table 5**; no inflation of signal was observed for the least abundant plasma cells).

Cell type	Number of tested *trans*-eQTLs	Lambda inflation	Nr. significant (FDR < 0.05)
CD4+ T-Cells	27582	1.14	4
CD8+ T-Cells	26216	1.14	4
Dendritic cells	4125	1.09	0
Plasma cells	1917	1.00	0
Classical monocytes	9211	1.12	0
Non-classical monocytes	5981	1.04	4
B cells	19319	1.10	0
NK cells	21122	1.17	25

Lambda inflation and replication numbers for *trans*-eQTLs per scRNA-seq cell type.

We subsequently attempted replication of *trans*-eQTLs per cell-type: we observed that we could significantly replicate 35 *trans*-eQTLs after correction for multiple testing (i.e. using a false discovery rate; FDR<0.05). The table above and **Supplementary Table 5** sums to 37 *trans*-eQTLs, which is due to the fact that 2 *trans*-eQTLs significantly replicated in two different cell-types. Except for one *trans*-eQTL, the allelic direction of all other significantly replicating *trans*-eQTLs was consistent with the allelic direction that was observed in the discovery dataset (97% concordance), providing an additional measure of support that these *trans*-eQTLs replicate.

Since single-cell data is very noisy, we combined the summary statistics of the 8 cell-types to increase statistical power to replicate *trans*-eQTLs. Since we had replication Z-scores on at most

8 cell-types for each *trans*-eQTL, we combined this information by averaging the Z-scores per *trans*-eQTL over the different cell-types. We then selected those *trans*-eQTLs with an absolute average replication Z-score of at least 1.96 (which corresponds to a nominal P<0.05). This resulted in the replication of 729 *trans*-eQTLs (**Supplementary Table 4**). Although the significance threshold selected here is somewhat arbitrary, we reasoned that, among this list of replicating *trans*-eQTLs, the allelic concordance with the discovery dataset should be high. This indeed was the case: for 614 of these *trans*-eQTLs (84.2%), the averaged replication Z-score had the same allelic direction as the discovery dataset, which is much higher than what is expected by chance (i.e. expected allelic concordance by chance is 50%, two-sided binomial test $P=3.63\times10^{-83}$).

These results indicate that a substantial fraction of the *trans*-eQTLs discovered in eQTLGen replicate with the same effect direction in single-cell eQTL data. We therefore conclude that at least these *trans*-eQTLs are not caused by cell-type composition differences, but instead reflect intracellular regulatory effects and that, due to the limited statistical power of the single-cell replication data, it is likely that this also holds true for many other *trans*-eQTLs.

Among the 729 replicating *trans*-eQTLs, we observed a strong enrichment for genes involved in cytokine-mediated signaling (hypergeometric test from ToppGene[126]; $P=3.3\times10^{-12}$, Benjamini-Hochberg FDR<0.05), with effects on *AIM2, ANXA1, BCL2, CD300LF, CD36, CD74, CISH, CSF3R, CXCL8, CXCR1, CXCR2, CXCR5, CXCR6, EDAR, FCGR1A, FOS, FYN, GBP1, H3C1, HLA-B, HLA-DQA1, HMOX1, IFIT1, IL18RAP, IL1B, IL2RB, IL4R, IL6ST, LTA, MX1, OAS3, PIK3R1, PPBP, STAT1, TNFRSF13B* and *TNFRSF14*. The SNPs that affect these genes are enriched for immune-mediated diseases and SNPs that affect cell-type composition.

Consortium authors

BIOS Consortium (Biobank-based Integrative Omics Study) – Author information

Management Team Bastiaan T. Heijmans (chair)[1], Peter A.C. 't Hoen[2], Joyce van Meurs[3], Aaron Isaacs[4], Rick Jansen[5], Lude Franke[6].
Cohort collection Dorret I. Boomsma[7], René Pool[7], Jenny van Dongen[7], Jouke J. Hottenga[7] (Netherlands Twin Register); Marleen MJ van Greevenbroek[8], Coen D.A. Stehouwer[8], Carla J.H. van der Kallen[8], Casper G. Schalkwijk[8] (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga[6], Lude Franke[6], Sasha Zhernakova[6], Ettje F. Tigchelaar[6] (LifeLines Deep); P. Eline Slagboom[1], Marian Beekman[1], Joris Deelen[1], Diana van Heemst[9] (Leiden Longevity Study); Jan H. Veldink[10], Leonard H. van den Berg[10] (Prospective ALS Study Netherlands); Cornelia M. van Duijn[4], Bert A. Hofman[11], Aaron Isaacs[4], André G. Uitterlinden[3] (Rotterdam Study).

Data Generation Joyce van Meurs (Chair)[3], P. Mila Jhamai[3], Michael Verbiest[3], H. Eka D. Suchiman[1], Marijn Verkerk[3], Ruud van der Breggen[1], Jeroen van Rooij[3], Nico Lakenberg[1]. Data management and computational infrastructure Hailiang Mei (Chair)[12], Maarten van Iterson[1], Michiel van Galen[2], Jan Bot[13], Dasha V. Zhernakova[6], Rick Jansen[5], Peter van 't Hof[12], Patrick Deelen[6], Irene Nooren[13], Peter A.C. 't Hoen[2], Bastiaan T. Heijmans[1], Matthijs Moed[1].

Data Analysis Group Lude Franke (Co-Chair)[6], Martijn Vermaat[2], Dasha V. Zhernakova[6], René Luijk[1], Marc Jan Bonder[6], Maarten van Iterson[1], Patrick Deelen[6], Freerk van Dijk[14], Michiel van Galen[2], Wibowo Arindrarto[12], Szymon M. Kielbasa[15], Morris A. Swertz[14], Erik W. van Zwet[15], Rick Jansen[5], Peter-Bram 't Hoen (Co-Chair)[2], Bastiaan T. Heijmans (Co-Chair)[1].

1. Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
2. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
3. Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands
4. Department of Genetic Epidemiology, Erasmus MC, Rotterdam, The Netherlands
5. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
6. Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands
7. Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
8. Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands
9. Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands
10. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands
11. Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
12. Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
13. SURFsara, Amsterdam, the Netherlands
14. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

15. Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

i2QTL Consortium – Author information

Marc Jan Bonder[1,2,3], Craig Smail[4,5], Michael J. Gloudemans[4], Laure Frésard[6], David Jakubosky[7,8], Matteo D'Antonio[9], Xin Li[6,10], Nicole M. Ferraro[4], Ivan Carcamo-Orive[11], Bogdan Mirauta[1], Daniel D. Seaton[1], Na Cai[1,12], Helena Kilpinen[13,14,15,16], Danilo Horta[1], Joshua W. Knowles[11], Erin N. Smith[17], Kelly A. Frazer[9,17], Stephen B. Montgomery[6], Oliver Stegle[1,2,3,12]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, CB10 1SD Cambridge, UK
2. European Molecular Biology Laboratory, Genome Biology Unit, 69117 Heidelberg, Germany
3. Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
4. Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA
5. Genomic Medicine Center, Children's Mercy Research Institute and Children's Mercy Kansas City, Kansas City, MO 64108, USA
6. Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
7. Biomedical Sciences Graduate Program, University of California, San Diego, La Jolla, CA 92093, USA
8. Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093, USA
9. Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA
10. Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, China
11. Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
12. Wellcome Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA Cambridge, UK
13. Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland
14. Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland
15. UCL Great Ormond Street Institute of Child Health, University College London, London, UK
16. Faculty of Medicine, Imperial College London, London, UK
18. Department of Pediatrics and Rady Children's Hospital, University of California, San Diego, La Jolla, CA, 92093, USA

Supplementary References

1. Hyde, C. L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* **48**, 1031–1036 (2016).

2. Dastani, Z. *et al.* Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals. *PLOS Genetics* **8**, e1002607 (2012).

3. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333–338 (2011).

4. Peden, J. *et al.* A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature genetics* **43**, 339–344 (2011).

5. 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 Genetics* **7**, e1002193 (2011).

6. Mozaffarian, D. *et al.* Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *The American Journal of Clinical Nutrition* **101**, 398–406 (2015).

7. Wu, J. H. Y. *et al.* Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: Results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circulation: Cardiovascular Genetics* **6**, 171–183 (2013).

8. Guan, W. *et al.* Genome-wide association study of plasma n6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circulation: Cardiovascular Genetics* **7**, 321–331 (2014).

9. Teumer, A. *et al.* Genome-wide association studies identify genetic loci associated with Albuminuria in diabetes. *Diabetes* **65**, 803–817 (2016).

10. Pattaro, C. *et al.* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nature Communications* **7**, 10023 (2016).

11. Cai, N. *et al.* Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).

12. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* **46**, 234–244 (2014).

13. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981–990 (2012).

14. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics* **47**, 1449–1456 (2015).

15. Benke, K. S. *et al.* A genome-wide association meta-analysis of preschool internalizing problems. *Journal of the American Academy of Child and Adolescent Psychiatry* **53**, 667-676.e7 (2014).

16. Bradfield, J. P. *et al.* A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature Genetics* **44**, 526–531 (2012).

17. van der Valk, R. J. P. *et al.* A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Human Molecular Genetics* **24**, 1155–1168 (2015).

18. Rob Taal, H. *et al.* Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nature Genetics* **44**, 532–538 (2012).

19. Felix, J. F. *et al.* Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Human Molecular Genetics* **25**, 389–403 (2016).

20. Cousminer, D. L. *et al.* Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Human Molecular Genetics* **22**, 2735–2747 (2013).

21. Cousminer, D. L. *et al.* Genome-wide association study of sexual maturation in males and females highlights a role for body mass and menarche loci in male puberty. *Human Molecular Genetics* **23**, 4452–4464 (2014).

22. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).

23. Horikoshi, M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nature Genetics* **45**, 76–82 (2013).

24. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine* **363**, 1211–1221 (2010).

25. Zheng, H. F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).

26. Berndt, S. I. *et al.* Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics* **45**, 501–512 (2013).

27. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

28. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

29. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics* **46**, 1173–1186 (2014).

30. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1285 (2013).

31. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

32. Van Den Berg, S. M. *et al.* Harmonization of neuroticism and extraversion phenotypes across inventories and cohorts in the Genetics of Personality Consortium: An application of item response theory. *Behavior Genetics* **44**, 295–313 (2014).

33. De Moor, M. H. M. *et al.* Meta-analysis of genome-wide association studies for personality. *Molecular Psychiatry* **17**, 337–349 (2012).

34. De Moor, M. H. M. *et al.* Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* **72**, 642–650 (2015).

35. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nature Genetics* **45**, 145–154 (2013).

36. van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).

37. Den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature Genetics* **45**, 621–631 (2013).

38. Lu, Y. *et al.* New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature Communications* **7**, 10495 (2016).

39. Kilpeläinen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nature Communications* **7**, 10494 (2016).

40. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452–1458 (2013).

41. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **47**, 979–986 (2015).

42. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* **43**, 246–252 (2011).

43. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics* **45**, 1353–1362 (2013).

44. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).

45. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* **47**, 1457–1464 (2015).

46. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature Communications* **6**, 8019 (2015).

47. Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature Genetics* **44**, 1137–1141 (2012).

48. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics* **43**, 1193–1201 (2011).

49. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics* **42**, 295–302 (2010).

50. Faraco, J. *et al.* ImmunoChip Study Implicates Antigen Presentation to T Cells in Narcolepsy. *PLoS Genetics* **9**, e1003270 (2013).

51. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Genetics* **45**, 664–669 (2013).

52. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics* **42**, 508–514 (2010).

53. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

54. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics* **47**, 381–386 (2015).

55. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics* **44**, 1341–1348 (2012).

56. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics* **42**, 105–116 (2010).

57. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics* **44**, 659–669 (2012).

58. Saxena, R. *et al.* Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nature Genetics* **42**, 142–148 (2010).

59. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229–3239 (2010).

60. Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60**, 2624–2634 (2011).

61. Walford, G. A. *et al.* Genome-wide association study of the modified Stumvoll Insulin Sensitivity Index identifies BCL2 and FAM19A2 as novel insulin sensitivity loci. *Diabetes* **65**, 3200–3211 (2016).

62. Prokopenko, I. *et al.* A Central Role for GRB10 in Regulation of Islet Function in Man. *PLoS Genetics* **10**, e1004235 (2014).

63. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics* **44**, 991–1005 (2012).

64. Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications* **7**, 11122 (2016).

65. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nature Genetics* **46**, 543–550 (2014).

66. Smoller, J. W. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: A genome-wide analysis. *The Lancet* **381**, 1371–1379 (2013).

67. Sullivan, P. F. *et al.* A mega-analysis of genome-wide association studies for major depressive disorder. *Molecular Psychiatry* **18**, 497–511 (2013).

68. Sklar, P. *et al.* Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* **43**, 977–985 (2011).

69. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45**, 1150–1159 (2013).

70. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Molecular Psychiatry* **19**, 1017–1024 (2014).

71. Ruderfer, D. M. *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715.e16 (2018).

72. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

73. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nature Genetics* **48**, 1043–1048 (2016).

74. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).

75. Day, F. R. *et al.* Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nature Genetics* **47**, 1294–1303 (2015).

76. Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Molecular Psychiatry* **19**, 253–258 (2014).

77. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624–633 (2016).

78. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).

79. Rietveld, C. A. *et al.* Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13790–13794 (2014).

80. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

81. Furberg, H. *et al.* Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics* **42**, 441–447 (2010).

82. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).

83. Orrù, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell* **155**, 242–256 (2013).

84. Roederer, M. *et al.* The genetic architecture of the human immune system: A bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).

85. Eicher, J. D. *et al.* Platelet-Related Variants Identified by Exomechip Meta-analysis in 157,293 Individuals. *American Journal of Human Genetics* **99**, 40–55 (2016).

86. Tajuddin, S. M. *et al.* Large-Scale Exome-wide Association Analysis Identifies Loci for White Blood Cell Traits and Pleiotropy with Immune-Mediated Diseases. *American Journal of Human Genetics* **99**, 22–39 (2016).

87. Chami, N. *et al.* Exome Genotyping Identifies Pleiotropic Variants Associated with Red Blood Cell Traits. *American Journal of Human Genetics* **99**, 8–21 (2016).

88. Leslie, R., O'Donnell, C. J. & Johnson, A. D. GRASP: Analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185–i194 (2014).

89. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369–375 (2012).

90. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

91. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* **49**, 139–145 (2017).

92. Zaykin, D. V. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology* **24**, 1836–1841 (2011).

93. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

94. Aguirre-Gamboa, R. *et al.* Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinformatics* **21**, 243 (2020).

95. Netea, M. G. *et al.* Understanding human immune function using the resources from the Human Functional Genomics Project. *Nature Medicine* **22**, 831–833 (2016).

96. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493–497 (2018).

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

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

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

100. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306**, 636–640 (2004).

101. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

102. Matys, V. *et al.* TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108–D110 (2006).

103. Matys, V. *et al.* TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**, 374–378 (2003).

104. Mathelier, A. *et al.* JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* **42**, D142–D147 (2014).

105. Friedman, R. C., Farh, K. K. H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* **19**, 92–105 (2009).

106. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).

107. Hsu, S. D. *et al.* MiRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research* **39**, (2011).

108. Griffiths-Jones, S. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144 (2006).

109. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

110. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database : the journal of biological databases and curation* **2016**, baw105 (2016).

111. Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nature Methods* **14**, 61–64 (2016).

112. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Computational Biology* **12**, 1–20 (2016).

113. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature Methods* **13**, 366–370 (2016).

114. 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).

115. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J Numer Anal* **33**, 1029–1047 (2013).

116. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics* **45**, 1238–1243 (2013).

117. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

118. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *The American Journal of Human Genetics* **100**, 571–580 (2017).

119. 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. *Human Molecular Genetics* **24**, 4746–4763 (2015).

120. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics* **49**, 131–138 (2017).

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

122. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

123. Van Der Wijst, M. G. P., De Vries, D. H., Brugge, H., Westra, H. J. & Franke, L. An integrative approach for building personalized gene regulatory networks for precision medicine. *Genome Medicine* **10**, 96 (2018).

124. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature Communications* **9**, 2282 (2018).

125. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research* **37**, W305-11 (2009).

126. 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).

127. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).

128. Myers, R. M. *et al.* A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biology* **9**, (2011).

129. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): A coordinate repressor of multiple neuron-specific genes. *Science* **267**, 1360–1363 (1995).

130. Chong, J. A. *et al.* REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).

131. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).

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

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

134. Deelen, P. *et al.* Improving the diagnostic yield of exome- sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nature Communications* **10**, 2837 (2019).

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

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

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

138. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).

139. Wang, J. H. *et al.* Aiolos regulates B cell activation and maturation to effector state. *Immunity* **9**, 543–553 (1998).

140. Dibner, C., Schibler, U. & Albrecht, U. The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks. *Annual Review of Physiology* **72**, 517–549 (2010).

141. Gachon, F., Olela, F. F., Schaad, O., Descombes, P. & Schibler, U. The circadian PAR-domain basic leucine zipper transcription factors DBP, TEF, and HLF modulate basal and inducible xenobiotic detoxification. *Cell Metabolism* **4**, 25–36 (2006).

142. Bass, J. & Lazar, M. A. Circadian time signatures of fitness and disease. *Science* **354**, 994–999 (2016).

143. Song, C. *et al.* REV-ERB agonism suppresses osteoclastogenesis and prevents ovariectomy-induced bone loss partially via FABP4 upregulation. *FASEB Journal* **32**, 3215–3228 (2018).

144. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**, (2020).

145. Kõks, G. *et al.* Smoking-induced expression of the GPR15 gene indicates its potential role in chronic inflammatory pathologies. *American Journal of Pathology* **185**, 2898–2906 (2015).

146. van Iterson, M. *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology* **18**, 19 (2017).

147. 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 Pulmonary Medicine* **17**, (2017).

148. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics* **50**, 956–967 (2018).

149. 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).

150. Sousa, I. *et al.* Polymorphisms in leucine-rich repeat genes are associated with autism spectrum disorder susceptibility in populations of European ancestry. *Molecular Autism* **1**, 7 (2010).

151. Agarwal, A. *et al.* Dysregulated expression of neuregulin-1 by cortical pyramidal neurons disrupts synaptic plasticity. *Cell Reports* **8**, 1130–1145 (2014).

152. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* **47**, 115–125 (2015).

153. Baechler, E. C. *et al.* Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 2610–2615 (2003).

154. Bennett, L. *et al.* Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *Journal of Experimental Medicine* **197**, 711–723 (2003).

155. 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 Biology* **12**, e1001759 (2014).

156. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

157. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* **7**, 500–507 (2012).