nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0089-9

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

Monique G. P. van der Wijst[1], Harm Brugge[1,3], Dylan H. de Vries[1,3], Patrick Deelen[1], Morris A. Swertz[1], LifeLines Cohort Study[2], BIOS Consortium[2] and Lude Franke[1*]

Genome-wide association studies have identified thousands of genetic variants that are associated with disease[1]. Most of these variants have small effect sizes, but their downstream expression effects, so-called expression quantitative trait loci (eQTLs), are often large[2] and celltype-specific[3–5]. To identify these celltype-specific eQTLs using an unbiased approach, we used single-cell RNA sequencing to generate expression profiles of ~25,000 peripheral blood mononuclear cells from 45 donors. We identified previously reported cis-eQTLs, but also identified new celltype-specific cis-eQTLs. Finally, we generated personalized co-expression networks and identified genetic variants that significantly alter co-expression relationships (which we termed 'co-expression QTLs'). Single-cell eQTL analysis thus allows for the identification of genetic variants that impact regulatory networks.

Previously, purified cell types[4,6–8] or deconvolution methods[9,10] have been used to identify celltype-specific eQTLs. However, these methods are biased toward specific cell types or are of limited use for less-abundant cell types and dependent on accurately defined marker genes[11]. In contrast, single-cell RNA sequencing (scRNA-seq) can be used to investigate rare cell types[12] and thus enables identification of celltype-specific eQTLs using an unbiased approach. Indeed, proof of concept was previously shown in a study on 15 individuals, where 92 genes were studied in 1,440 cells[13].

Here we studied celltype-specific effects of genetic variation on genome-wide gene expression by generating scRNA-seq data of ~25,000 peripheral blood mononuclear cells (PBMCs) from 45 donors of the population-based cohort study Lifelines Deep[14]. After quality control (Methods and Supplementary Fig. 1), we first assessed the extent to which previously reported cis-eQTLs from bulk whole blood, using either 94 DeepSAGE samples[15] (a 3′-end-oriented RNA-sequencing strategy similar to our scRNA-seq approach) or 2,116 RNA-seq[11] samples, also show significant effects in the scRNA-seq dataset. For this analysis, we treated the scRNA-seq data as representing bulk PBMCs (by averaging expression levels of all cells per gene per sample, referred to as 'bulk-like PBMCs'). We detected 50 and 311 significant cis-eQTLs (gene-level false-discovery rate (FDR) of 0.05) that were previously reported in the DeepSAGE[15] and RNA-seq[11] study, respectively (Fig. 1a and Supplementary Table 1). Although only small proportions (8% and 1%, respectively) of previously reported cis-eQTLs were significant in our scRNA-seq analysis, 96% and 90.4% had identical allelic directions as in the DeepSAGE[15] and RNA-seq[11] studies, respectively, indicating that these cis-eQTLs reflect similar regulatory effects. The few discordant eQTLs may reflect the slightly different sample composition of both datasets (PBMCs versus whole blood) and the relatively few sequence reads targeting the 3′-end of genes in the bulk RNA-seq dataset.

We subsequently performed a genome-wide cis-eQTL discovery analysis on the bulk-like PBMCs. Separate cis-eQTL analyses were conducted on each of the identified major cell types (cell type classification was performed using Seurat[16]; Supplementary Fig. 2a,b) by averaging the normalized gene expression of all cells per cell type, gene, and donor. In total, 379 unique top cis-eQTLs were identified, reflecting 287 unique eQTL genes (gene-level FDR of 0.05; Table 1), as sometimes, in different cell types, different single-nucleotide polymorphisms (SNPs) showed the most significant association for an eQTL gene. While 331 (reflecting 249 unique cis-eQTL genes) of these 379 cis-eQTLs were significant in the bulk-like PBMC eQTL analysis, 48 cis-eQTLs (reflecting 38 unique cis-eQTL genes) were only detected in specific cell types (i.e., 'celltype-dependent' eQTLs; Supplementary Table 2).

We subsequently attempted to replicate these eQTLs. For the 249 eQTL genes found in the bulk-like PBMC analysis, 233 cis-eQTLs were testable and 181 (78%) were associated with the same SNP (90.1% shared allelic direction; Supplementary Table 2) in the wholeblood RNA-seq eQTL dataset[11]. For the 48 celltype-dependent cis-eQTLs, 29 (60%) were replicated in the RNA-seq dataset[11]. This lower percentage suggests that in bulk RNA-seq datasets, celltype-dependent eQTLs might become too diluted, resulting in low statistical power to recover these. While this most likely happens for rare cell types, we also observed this in common cell types. For instance, in the most abundant cell type (CD4$^+$ T cells), rs2272245 significantly affects expression of the *TSPAN13* gene in cis ($P = 2.21 \times 10^{-6}$). However, this effect was not significant in the bulk-like PBMCs ($P = 0.88$), because *TSPAN13* is lowly expressed in CD4$^+$ T cells, whereas it is highly expressed in dendritic cells, where it did not show a cis-eQTL effect (Fig. 1b). Cis-eQTLs might also be missed in bulk data, because they might show opposite allelic effects across different cell types. We could not study this in detail due to lack of power, given the sample size and limited number of cells for rare cell types (Supplementary Fig. 2c). Nevertheless, in CD4$^+$ T cells, the A allele of rs4804315 significantly decreased expression of *ZNF414* in cis ($P = 6.09 \times 10^{-6}$), whereas in natural killer cells this allele increased expression of *ZNF414* at nominal significance ($P = 0.0339$; Fig. 1b). However, the possibility that, specifically in natural killer cells, the effect of rs4804315 on *ZNF414* expression is the result of a residual effect on *ZNF414* expression of a second, independent variant cannot be excluded.

Since some cis-eQTLs did not replicate in the wholeblood bulk RNA-seq data, we subsequently investigated eQTL datasets of purified cell types. Indeed, 3 of 19 remaining celltype-dependent

[1]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [2]A list of members and affiliations appears in the Supplementary Note. [3]These authors contributed equally: Harm Brugge, Dylan H. de Vries. *e-mail: lude@ludesign.nl

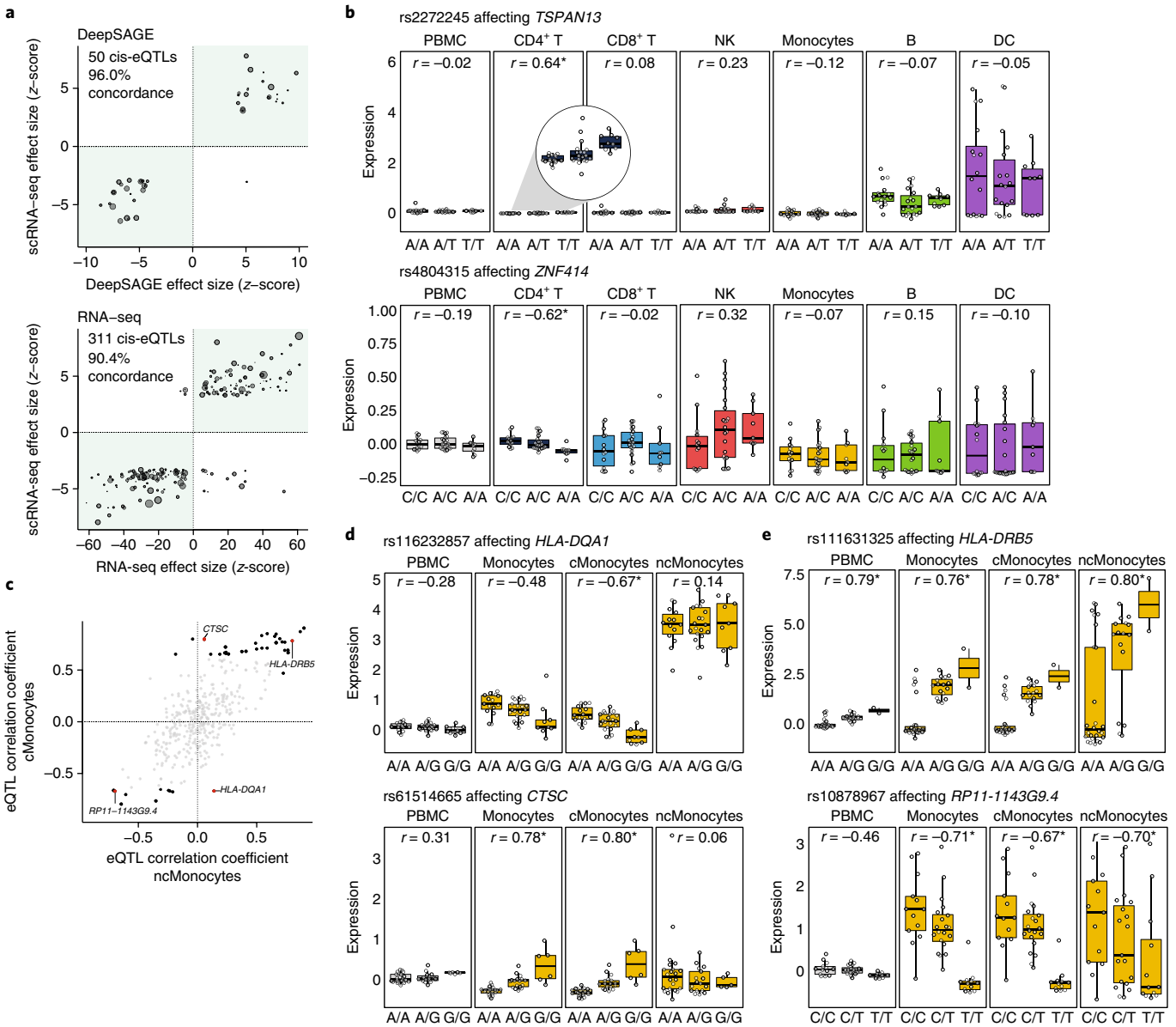

Fig. 1 | Cis-eQTL analysis in single-cell RNA-seq data. a, Effect size of the cis-eQTLs detected in the bulk-like PBMC scRNA-seq sample in which the analysis was confined to previously reported cis-eQTLs in (top) wholeblood DeepSAGE or (bottom) bulk RNA-seq data. The numbers and percentages represent, respectively, the detected cis-eQTLs and their concordance (i.e., same allelic direction, blue quadrants) between the bulk-like PBMC population scRNA-seq eQTLs and (top) the whole-blood DeepSAGE or (bottom) bulk RNA-seq data. The size of each dot represents the mean expression of the cis-regulated gene in the total scRNA-seq dataset. **b**, Examples of undetectable cis-eQTLs in the bulk-like PBMC population caused by (top) masking of the cis-eQTL present in CD4$^+$ T cells but absent in dendritic cells (DCs) with comparatively high expression of the cis-regulated gene or (bottom) opposite allelic effects in CD4$^+$ T and natural killer (NK) cells. **c**, Spearman's rank correlation coefficient for cMonocytes against ncMonocytes of all top eQTLs that were identified in the total dataset or in at least one (sub)cell cluster (see Supplementary Table 2). Significant correlations are shown in black (four red highlighted examples are shown in **d** and **e**); nonsignificant correlations are in gray. **d**, Cis-eQTLs specifically affecting expression in cMonocytes and not ncMonocytes. **e**, Cis-eQTLs significantly affecting expression in both cMonocytes and ncMonocytes. Each dot represents the mean expression of the eQTL gene in a donor. Box plots show the median, first and third quartiles, and 1.5 × the interquartile range. *r*, Spearman's rank correlation coefficient; *FDR ≤ 0.05.

cis-eQTLs were detected (each with consistent allelic direction) in purified eQTL datasets of the Blueprint consortium (naive CD4$^+$ T cells and CD14$^+$ monocytes)[17] or Kasela et al. (CD4$^+$ and CD8$^+$ T cells)[6] (Supplementary Table 3). Hence, only 16 celltype-dependent cis-eQTLs were not identified before using bulk eQTL datasets of blood or purified immune cells. Although some cis-eQTLs were only significant in specific cell types, this does not prove celltype-specificity; power is lacking to detect many cis-eQTLs, particularly in less-abundant cell types. Ways to partially overcome this include using methods that consider multiple eQTL datasets together, such as eQTL-BMA[18] or Meta-Tissue[19]. However, these methods are currently computationally too demanding for large scRNA-seq data or do not define the cell type in which the eQTL effect occurs[19,20].

A major advantage of using scRNA-seq data is its flexibility, which allows any cell population of interest to be selected for eQTL analysis. In contrast, when using RNA-seq data of purified cell types, one cannot retrieve data from cell subtypes. Moreover, while finer differences between cell subtypes may be detectable using gene

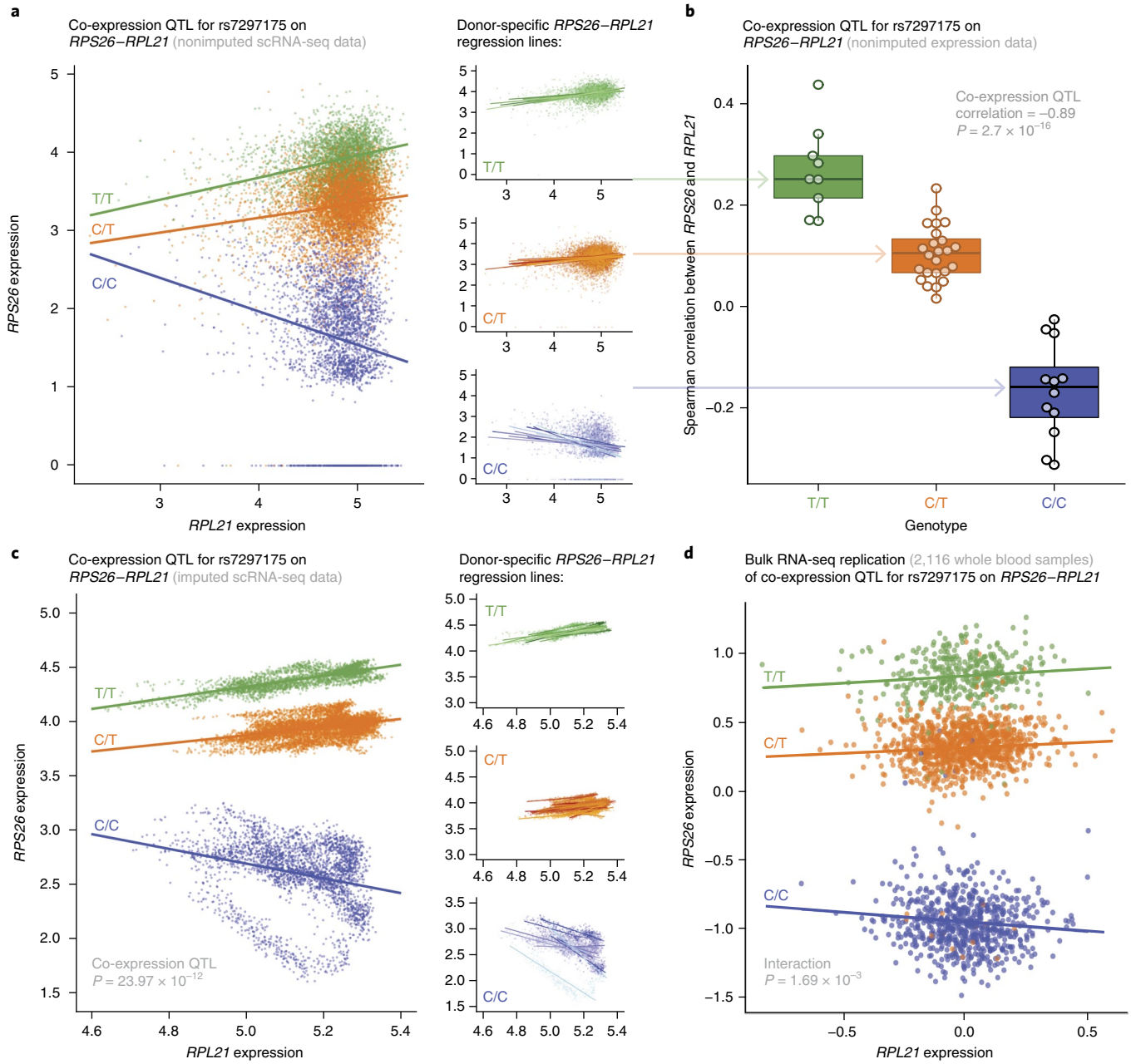

Fig. 2 | Most significant co-expression QTL in the CD4⁺ T cells. a, The nonimputed expression of *RPS26* and *RPL21* of all individual CD4⁺ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panels, respectively. Each data point represents a single cell. The nominal *P* value is given for the co-expression QTL. **b**, The Spearman's rank correlation coefficient (*r*) between *RPS26* and *RPL21* expression, stratified by SNP rs7297175 genotype in the CD4⁺ T cells per donor. Each data point represents a single donor. Box plots show the median, first and third quartiles, and 1.5 × the interquartile range. The nominal *P* value is given for the co-expression QTL. **c**, The imputed expression of *RPS26* and *RPL21* of all individual CD4⁺ T cells colored by genotype (left panel) and stratified per SNP rs7297175 genotype (right panels). Genotype- and donor-specific regression lines are shown in the left and right panel, respectively. Each data point represents a single cell. **d**, The expression of *RPS26* and *RPL21* in wholeblood bulk RNA-seq samples, colored by SNP rs7297175 genotype. Genotype-specific regression lines are shown. Each data point represents a single bulk RNA-seq sample. The nominal *P* value is given for the interaction effect.

expression profiles, it is not always recapitulated by different cell membrane markers, complicating cell sorting. Here we show the added value of performing eQTL analysis on cell subtypes using two monocyte subsets: classical (cMonocytes) and nonclassical monocytes (ncMonocytes). When plotting Spearman's rank correlation of each top eQTL for cMonocytes against that for ncMonocytes, several examples were identified that pinpointed the eQTL effect specifically to cMonocytes (Fig. 1c). Two such examples, which were previously identified in RNA-seq data of purified CD14⁺ monocytes[17], are shown in Fig. 1d. The scRNA-seq data now allowed us to specifically assign these effects to cMonocytes (Fig. 1d). Despite having lower power for detecting eQTLs in ncMonocytes due to an abundance almost five times lower compared to cMonocytes (Supplementary Fig. 2b), power in the ncMonocytes remains sufficiently high to detect several other significant ncMonocyte cis-eQTLs (Fig. 1e and Supplementary Table 2).

Another opportunity of scRNA-seq data is to use it for determining whether genetic variants can alter gene co-expression. Although

Table 1 | Cis-eQTL genes identified per cell type

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

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

recently genes and environmental factors altering the effect size of eQTLs ('context-specific eQTLs') have been identified in bulk RNA-seq eQTL datasets[11,21], a large sample size was required to ensure sufficient power. In contrast, scRNA-seq data enables generation of co-expression networks on an individualdonor basis, which vastly reduces the number of samples required to identify SNPs altering co-expression relationships. This enabled us to study whether SNPs showing cis-eQTL effects also affect the co-expression relationship of the cis-eQTL genes with other genes, which we define as 'co-expression QTLs'. We confined our analysis to the most abundant cell type (CD4⁺ T cells) and calculated the co-expression between individual pairs of genes using Spearman's rank correlation. We restricted the analysis to the 145 cis-eQTL genes identified in CD4⁺ T cells (Table 1), thereby increasing the likelihood of finding co-expressed genes that are modulated by the same genetic variant. Of these, 102 genes showed variance in gene expression within each of the 45 donors and were investigated. For two of these genes, we identified significant co-expression QTLs: 93 co-expression QTLs were detected for *RPS26* and one for *HLA-B* ($P \leq 1.27 \times 10^{-7}$, corresponding to an eQTL-gene level FDR of 0.05). The most significant interaction was found for rs7297175, affecting the co-expression between *RPS26* and *RPL21* ($P = 2.70 \times 10^{-16}$; Fig. 2a,b). When using a more liberal FDR of 0.10 ($P \leq 4.72 \times 10^{-7}$), we identified significant co-expression QTLs for three eQTL genes (Supplementary Table 4): 13 additional co-expression QTLs were found for *RPS26* and one for *SMDT1*. Due to co-expression between genes, we cannot rule out that the 106 co-expression QTLs identified for *RPS26* are actually representing just one effect.

To assess the robustness of the identified co-expression QTLs, we tested whether they remained significant after geneexpression imputation, which was used to overcome the problem that, in scRNA-seq data, many genes are often undetected despite being expressed (i.e., zero-inflated expression). Several computational strategies have been developed to do this[22–24]. However, most current methods are either computationally too demanding for large datasets like ours[23] or cannot sufficiently impute the 94.1% zero-values present in our dataset[24]. To overcome this, we used MAGIC[22], a method that imputes geneexpression levels for nearly every gene. To prevent imputation from removing effects of genetic differences between donors or cell types, we performed imputation for each donor separately and again only for CD4⁺ T cells (see "Data availability" in Methods). In general, imputation worked well, but in some circumstances artifacts were introduced (Supplementary Fig. 3). Therefore, we only used the imputed geneexpression data to determine whether the co-expression QTLs identified before imputation remained significant after imputation (Supplementary Table 4). Of the three eQTL genes that were involved in a co-expression QTL, two of three top co-expression QTLs—rs7297175 (affecting the co-expression between *RPS26* and *RPL21*, $P = 3.97 \times 10^{-12}$; Fig. 2c) and rs4147641 (affecting the co-expression between *SMDT1* and *RPS3A*, $P = 2.57 \times 10^{-4}$)—remained after imputation (Supplementary Table 4). Subsequently, we were able to replicate both effects in a wholeblood bulk RNA-seq eQTL dataset[11] ($P = 1.69 \times 10^{-3}$ for *RPS26–RPL21* (Fig. 2d), $P = 1.59 \times 10^{-4}$ for *SMDT1–RPS3A*; Supplementary Table 4). Notably, SNP rs7297175, affecting the co-expression between *RPS26* and 106 other genes, is in near-perfect linkage disequilibrium with the type I diabetes SNP rs11171739[25] ($r^2 = 0.98$). Therefore, the numerous co-expression QTLs for *RPS26* may shed new light on *RPS26* and its link with type I diabetes. This interaction effect was also observed in other cell types (Supplementary Fig. 4), indicating that it is not celltype-specific. In addition, various analyses were performed to rule out potential technical confounders (Methods).

The co-expression QTL analysis as outlined above highlights another advantage of scRNA-seq data; with PBMCs from only 45 donors, we could identify effects that would otherwise only become apparent in large-scale (2,116 samples) bulk RNA-seq eQTL datasets[11]. Due to Simpson's paradox[26], it may occur that when looking at all individuals together, the interaction between two genes does not show a correlation, while each of the individuals separately do show a correlation. Therefore, even though the effect may be observed in bulk RNA-seq data, the true correlation will only be identified using scRNA-seq data.

The eQTL and co-expression QTL analyses performed in this study show the benefit of scRNA-seq data for linking genetic variation to geneexpression regulation. In addition to these analyses, we expect scRNA-seq data to offer many other opportunities for selecting cells of interest for eQTL and co-expression QTL analysis. For example, one could use the intercellular variation within scRNA-seq data to group cells along the cell cycle[13], along a differentiation path[27], or along a response to an environmental stimulus[28]. By doing so, one might identify eQTLs or co-expression QTLs that are influenced by cell cycle phase, differentiation, or environmental status.

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

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0089-9.

Received: 17 August 2017; Accepted: 23 February 2018;
Published online: 2 April 2018

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
3. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
4. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).

5. Fu, J. et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8**, e1002431 (2012).
6. Kasela, S. et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet.* **13**, e1006643 (2017).
7. Naranbhai, V. et al. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
8. Ishigaki, K. et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* **49**, 1120–1125 (2017).
9. Westra, H. J. et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).
10. Venet, D., Pecasse, F., Maenhaut, C. & Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics* **17** (Suppl. 1), S279–S287 (2001).
11. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
12. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573, https://doi.org/10.1126/science.aah4573 (2017).
13. Wills, Q. F. et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).
14. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
15. Zhernakova, D. V. et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
16. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
17. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
18. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
19. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491 (2013).
20. Duong, D. et al. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics* **33**, i67–i74 (2017).
21. Knowles, D. A. et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699–702 (2017).
22. van Dijk, D. et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. Preprint at *bioRxiv* https://doi.org/10.1101/111591 (2017).
23. Huang, M. et al. Gene expression recovery for single cell RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/138677 (2017).
24. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
26. Simpson, E. H. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. Ser. B Methodol.* **13**, 238–241 (1951).
27. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
28. Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).

Acknowledgements

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

Author contributions

M.G.P.v.d.W. generated the scRNA-seq data. M.G.P.v.d.W., H.B., and D.H.d.V. performed bioinformatics and statistical analyses. P.D. and the BIOS Consortium performed replication of co-expression QTLs. M.G.P.v.d.W. and L.F. designed the study and wrote the manuscript. M.A.S. and the LifeLines Cohort Study provided biomaterials, genotype data, and computational resources. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0089-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Ethics approval and consent to participate. 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 PBMCs. Whole blood of 47 donors from the general-population Lifelines Deep (LLD) cohort[14] was drawn into EDTA-vacutainers (BD). Within 2 h, peripheral blood mononuclear cells (PBMCs) were isolated using Cell Preparation Tubes with sodium heparin (BD). For all procedures, PBMCs were kept in RPMI1640 supplemented with 50 µg/mL gentamicin, 2 mM L-glutamine, and 1 mM pyruvate. Isolated PBMCs were cryopreserved in RPMI1640 containing 40% FCS and 10% DMSO. Within one month, PBMCs were further processed for scRNA-seq. First, cells were thawed in a 37 °C water bath until almost completely thawed, after which the cells were slowly washed in warm medium. After washing, cells were resuspended in medium and incubated for 1 h in a 5° slant rack at 37 °C in a 5% CO_2 incubator. After this 1 h resting period, cells were washed twice in medium supplemented with 0.04% bovine serum albumin. Cells were counted using a hemocytometer and cell viability was assessed by Trypan Blue. Eight sex-balanced sample pools were prepared, each containing 1,750 cells/donor from six (or five) donors (10,500 cells).

Single-cell library preparation and sequencing. Single cells were captured using the 10x Chromium controller (10x Genomics) according to the manufacturer's instructions (document CG00026) and as previously described[29]. Each sample pool was loaded into a different lane of a 10x chip (Single Cell A Chip Kit, 120236). cDNA libraries were generated using the Single Cell 3′ Library & Gel Bead kit version 2 (120237) and i7 Multiplex kit (120262), in line with the company's guidelines. These libraries were sequenced using a custom program (27-9-0-138) on eight lanes of an Illumina HiSeq4000 using a 75-bp paired-end kit, per GenomeScan (Leiden, The Netherlands) sequencing guidelines. In total, 28,855 cells were captured and sequenced to an average depth of 74 kb.

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

Demuxlet algorithm: demultiplexing samples per lane and doublet detection. Genotypes of the LLD samples were previously generated[14] and were phased using Eagle v2.3[30] and imputed with the HRC reference panel[31] using the Michigan Imputation Server[32]. As genotype data for each donor (except two) was available, we used the Demuxlet method[33], which uses variable SNPs between the pooled individuals to determine which cell belongs to which individual and to identify doublets (two cells encapsulated in a single droplet by the 10x Chromium controller).

To determine how well every genotype matches each cell, a likelihood score was calculated by the formula

$$L_c(s) = \prod_{v=1}^{V}\left[\sum_{g=0}^{2}\left\{\prod_{i=1}^{d_{cv}}\left(\sum_{e=0}^{1} Pr(b_{cvi}|g,e)\right)P_{sv}^{(g)}\right\}\right]$$

Here, c is the cell, s is the individual, v denotes the unique genetic variants (SNPs) found on the reads of the cell, and d_{cv} the number of unique reads overlapping with the vth variant from the cth cell. b_{cvi} is the variant-overlapping base call from the ith read, representing reference (R), alternate (A), and other (O) alleles respectively. e_{cvi} is a latent variable indicating whether the base call is correct (0) or not (1), and g is the true genotype. This likelihood score was calculated by taking into account the genotype probabilities of a sample at all known SNPs, the variant-overlapping base calls with base quality (Phred quality score) > 15, and a probability that the base was not called correctly, which is fixed at 0.001. In this way, for each pool of cells, the genotype within this pool with the highest likelihood was assigned as the most likely person the cell belonged to.

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

$$L_c(s_1,s_2,\alpha) = \prod_{v=1}^{V}\left[\sum_{g_1g_2}\left\{\prod_{i=1}^{d_{cv}}\left(\sum_{e=0}^{1}(1-\alpha)Pr(b_{cvi}|g_1,e)+\alpha\,\mathrm{Pr}(b_{cvi}|g_2,e)\right)P_{sv}^{(g_1)}P_{sv}^{(g_2)}\right\}\right]$$

Here, s_1 and s_2 are the two individuals, g_1 and g_2 the corresponding true genotypes and α is the expected proportion of the SNPs in every cell for each of the individuals. An α of 0.5 was consistently used, assuming a 50/50 ratio. The maximum likelihood in the mixed-genotype case was divided by the maximum likelihood in the singlet case to obtain a likelihood ratio. If this ratio was less than $1/t$ for some number t, the cell was considered a singlet of the sample corresponding to the maximum singlet likelihood. If the ratio was greater than t, the cell was considered a doublet. When the ratio was in between $1/t$ and t, the cell was called inconclusive: no confident call could be made from which sample(s) the cell originated. The decision boundary factor t was fixed at 2. In theory, if there are n samples in a lane, $(n-1)/n$ doublets can be identified using the Demuxlet algorithm, because doublets from the same individual ($1/n$) cannot be identified. Further details of the algorithm can be found in Kang et al.[33]

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

Two additional tests were performed to confirm the correct assignment of cells using Demuxlet. First, we determined what would happen if the cells did not match with their genotypes by taking six random genotypes not present in the sample pool itself. This resulted in 0.02% of the cells assigned as singlets, 0.03% inconclusive and 99.95% assigned as doublets. Second, the number of reads mapping to the Y-chromosome was determined for the singlets of each donor. Cells belonging to a female donor showed (almost) no Y-reads (mismapping reads[34] may explain the few sporadic Y-reads), whereas the majority of cells from male donors did (Supplementary Fig. 1b). So, the correct gender for each of the donors could be confirmed by looking at the number of Y-reads. These tests indicated that the Demuxlet method was correctly assigning cells to their respective donor and suitable for detecting sample swaps.

Cell type classification. Version 1.4 of the R package Seurat[16] was used to determine the cell types using the raw UMI counts from CellRanger. First, all genes that were not detected in ≥3 cells were removed. Cells in which > 5% of the UMIs mapped to the mitochondrially encoded genes were discarded, as this can be a marker of poor-quality cells; broken cells will leak cytoplasmic RNA, while the mitochondrial RNA content is retained inside the mitochondria[35]. Also, cells expressing > 3,500 genes were considered outliers and discarded (Supplementary Fig. 1c and Supplementary Table 6). Finally, all cells that were marked as doublets or inconclusive by the Demuxlet method were discarded. Supplementary Fig. 1d shows a t-distributed stochastic neighbor embedding (t-SNE) plot[36] in which all cells failing the above QCs are visualized. Library-size normalization was performed on the UMI-collapsed gene expression for each barcode by scaling the total number of transcripts per cell to 10,000. The data was then log2 transformed. In total, 25,291 cells and 19,723 genes (an average of 1,147 detected genes/cell; see "Data availability" section below) were used in the cell type determination.

Linear regression was used to regress out the total number of UMIs and the fraction of mitochondrial transcript content per cell. The variable genes were identified using Seurat's MeanVarPlot function, which sorts all genes into 20 bins based on their average expression (the mean of nonzero values) and calculates the dispersion (s.d. of all values) within each bin. Standard parameters were used, except that the bottom geneexpression cut-off (x.low.cutoff) and the bottom dispersion cut-off (y.cutoff) were each set to 1.0, resulting in the identification of 1,090 genes. These 1,090 variable genes were used in the principal component analysis (PCA). The first 16 principal components were used for cell clustering using Seurat's FindCluster function (default parameters, resolution 1.2) and a t-SNE plot was used to visualize this. Based on known marker genes and differentially expressed genes per cluster (found using Seurat's FindMarkers function), we could assign 11 cell types to the clusters, including some smaller cell subtypes (Supplementary Fig. 2a,b and Supplementary Table 7). The smallest cluster we could detect consisted of plasma cells, making up 0.3% of the total PBMC population.

eQTL analysis. To find the association between genotype and expression per cell type, genome-wide cis-eQTL analysis for 18,264 genes (only autosomal genes, gene expressed in at least 3 cells within the total dataset and in at least 1 cell within the cell type queried, within 100 kb distance of the SNP and the gene midpoint, MAF > 0.1, call rate > 0.95, Hardy–Weinberg equilibrium $P > 0.001$) was performed using our previously described eQTL pipeline, version 1.2.4 F (Supplementary Table 2 and see "Data availability")[11]. To assure sufficient power, cell types were merged to a more general classification: $CD4^+$ T cells, $CD8^+$ T cells, NK cells ($CD56^{dim}CD16^+$ and $CD56^{bright}CD16^{+/-}$), monocytes ($CD14^{bright}CD16^-$ cMonocyte and $CD14^{dim}CD16^+$ ncMonocyte), B cells and DCs ($CD1C^+$ myeloid, mDC, and plasmacytoid, pDC). The mean expression per gene per cell type per donor was calculated on the normalized (z-score transformed) expression and used as input for the eQTL analysis. eQTLs were mapped using Spearman's rank correlation coefficient on imputed genotype dosages. eQTLs were considered significant at

a gene-level FDR of 0.05. To control the FDR at 0.05 we used the permutation method described in our previous study[2]. Here we permute the link between the genotypes and expression data and create an overall null distribution using all genes. We performed, in total, 10 permutations and for each gene use the total null distribution of all genes to determine a gene-level FDR; during FDR estimation only the most significant SNP per gene is used, both for the real analysis and for each of the permutations.

Concordance and detection. Concordance with previously characterized, independent top eQTLs from a wholeblood DeepSAGE (3′-end transcriptomics)[15] and RNA-seq study[11] were computed. For this, the mean expression per gene per individual of all cells was calculated, and the cis-eQTL mapping was confined to the independent top eQTLs found in the DeepSAGE[15] or RNA-seq[11] studies. Subsequently, detection of the same SNP–gene combination and concordance (with same allelic direction) were assessed between the significant top effects (Supplementary Table 1). We also determined how many of the 379 top eQTLs in our scRNA-seq dataset could be detected and with which allelic direction within the wholeblood RNA-seq study[11]. Similarly, we assessed detection rates and concordances with two studies containing RNA-seq data of purified cell types: Kasela et al. performed eQTL analysis on purified $CD4^+$ and $CD8^+$ T cells[6], whereas the data from the Blueprint consortium contains purified $CD14^+$ monocytes and naive $CD4^+$ T cells[17] (Supplementary Tables 2 and 3). Moreover, for the eQTLs that were specifically detected in cMonocytes and not ncMonocytes (Fig. 2d), detection rates and concordances were determined using the RNA-seq data of the purified $CD14^+$ monocytes from the Blueprint consortium[17].

Single-cell gene expression imputation. To overcome the zero-inflated expression, the computational method MAGIC[22] was used to impute practically all values of genes with at least some expression. MAGIC imputation (using the following parameters: 20 PCs, $t=4$, $k=9$, $\mathrm{ka}=3$, $\varepsilon=1$) was performed separately per donor and only in the $CD4^+$ T cells (see "Data availability"). The effect of MAGIC imputation was validated by comparing the co-expression of typical celltype-specific marker genes (Supplementary Fig. 3).

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

To determine the number of cis-eQTL genes for which we had identified a significant co-expression QTL, we performed 100 permutations (see "Data availability"). For the real analysis, we denoted the most significant co-expression QTL *P* value for each of the tested 102 eQTL genes (Supplementary Table 4). For each permutation, we shuffled the genotype identifiers and re-ran the above analysis, including determination of the most significant co-expression QTL *P* value for each of the 102 eQTL genes (see "Data availability"). This subsequently enabled us to calculate an eQTL gene-level FDR[2] (using exactly the same multiple-testing correction procedures as we employed for the detection of cis-eQTLs; see "eQTL analysis" section). An eQTL gene-level FDR of 0.05 was considered significant, i.e., the *P* value threshold of the most significant co-expression QTL *P* values at which 5% of the co-expression QTLs are significant in the permuted compared to the real data.

All significant co-expression QTLs were discovered using nonimputed gene expression data. We then assessed whether these co-expression QTLs were also significant when using the MAGIC-imputed gene expression data. Subsequently, we tested whether these co-expression QTLs replicated using a large wholeblood bulk RNA-seq dataset[11] (Supplementary Table 4). Finally, we attempted to falsify the observed co-expression QTL for rs7297175 on the co-expression between *RPS26* and *RPL21* by checking the following potential confounders:

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

Accession codes. EGA: Processed (de-anonymized) single-cell RNA-seq data, EGAS00001002560.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

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

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

Code availability. The original R code for Seurat[16] (https://github.com/satijalab/seurat), Demuxlet[33] (https://github.com/statgen/demuxlet), MAGIC[22] (https://github.com/KrishnaswamyLab/magic) and our in-house eQTL pipeline[2] (https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline) can be found at GitHub. All custom-written code is made available via GitHub (https://github.com/molgenis/scRNA-seq).

References

29. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
30. Loh, P. R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
31. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
32. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
33. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
34. Rosser, Z. H., Balaresque, P. & Jobling, M. A. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. *Am. J. Hum. Genet.* **85**, 130–134 (2009).
35. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
36. van de Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

nature research

Corresponding author(s): L. Franke

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

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

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

Experimental design

1. Sample size

Describe how sample size was determined.

In this proof of concept study 47 donors were included. Previous eQTL studies have shown that this size is sufficient to detect eQTLs.

2. Data exclusions

Describe any data exclusions.

Two donors were included from downstream analysis as no genotype information was available to confidently assign the cells to these donors.

3. Replication

Describe whether the experimental findings were reliably reproduced.

77.7% (181/233) of the tested top cis-eQTLs (16/249 could not be tested as gene expression data was not available in the whole blood RNA-seq data) found in the total PBMCs were replicated (with 90.1% concordance) in whole blood RNA-seq data. The top co-expression QTLs found in CD4+ T cells remained after imputation and were replicated in whole blood RNA-seq data.

4. Randomization

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

The 47 donors were semi-randomly assigned to any of the 8 sample pools, taking into account that no relatives and approximately an equal number of male and females were included in each sample pool.

5. Blinding

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

All data was anonimyzed during data collection. During analysis, genotype information was used to assign each cell to its donor.

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

6. Statistical parameters

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

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

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

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	CellRanger version 1.3, R package Seurat version 1.4, Demuxlet (http://www.biorxiv.org/content/early/2017/05/15/118778), eQTL pipeline version 1.2.4F (https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline). All custom-made code is made available via GitHub (https://github.com/molgenis/systemsgenetics/scRNA-seq).

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

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Processed (deanonimyzed) single-cell RNA-seq data, including a text file that links each cell barcode to its respective donor, has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002560. Genotype data can be obtained and requested through our website (https://molgenis58.target.rug.nl/scrna-seq/) and will be made available through the Lifelines workspace.

9. Antibodies

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

10. Eukaryotic cell lines

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

Animals and human research participants

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

11. Description of research animals

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

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	All research participants were enrolled in the general population (Northern part of the Netherlands) cohort Lifelines Deep (http://bmjopen.bmj.com/content/5/8/e006772.long).