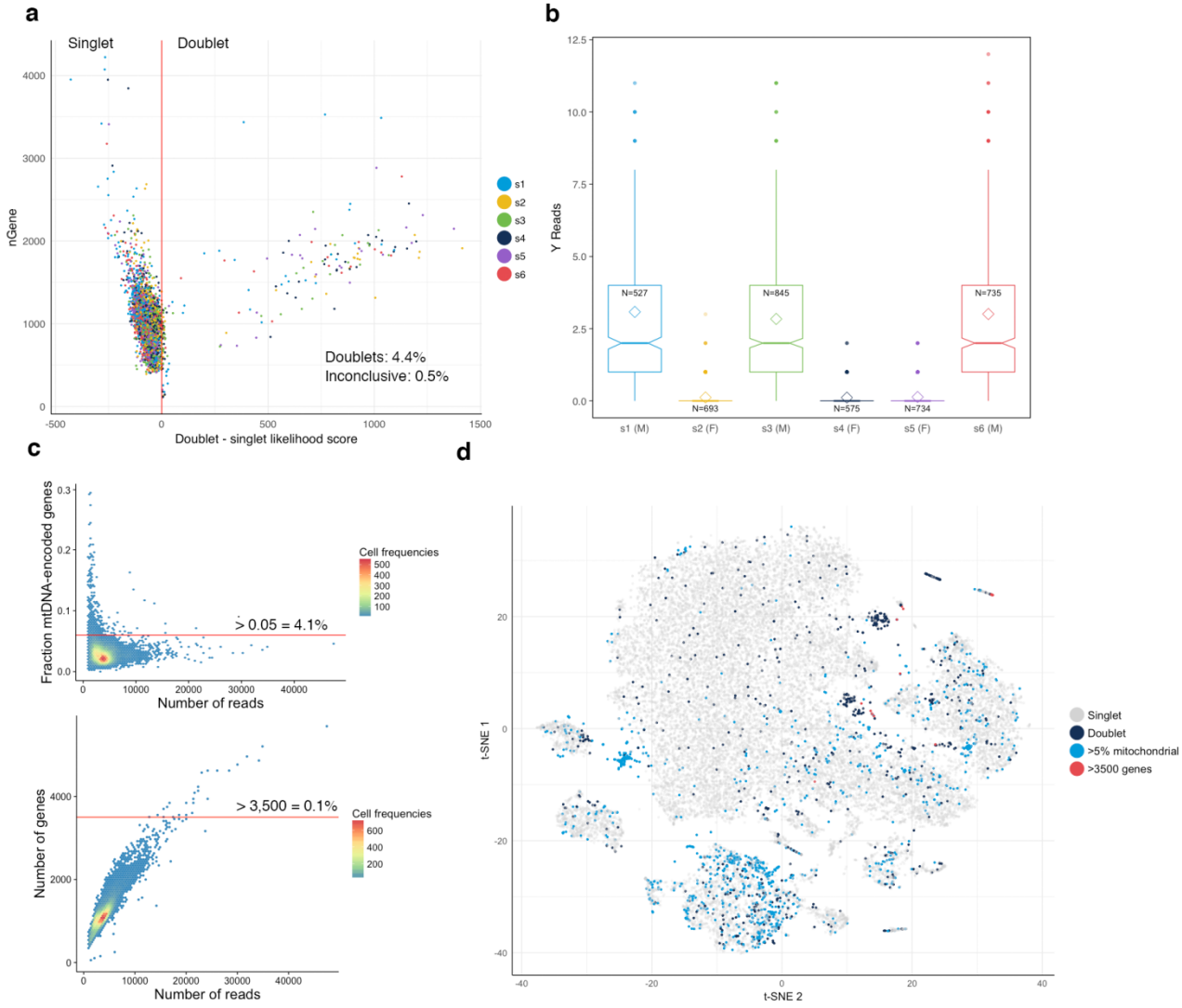
Supplementary information

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

Monique G.P. van der Wijst, Harm Brugge, Dylan H. de Vries, Patrick Deelen, Morris A. Swertz, Lifelines Cohort Study, BIOS consortium, Lude Franke.

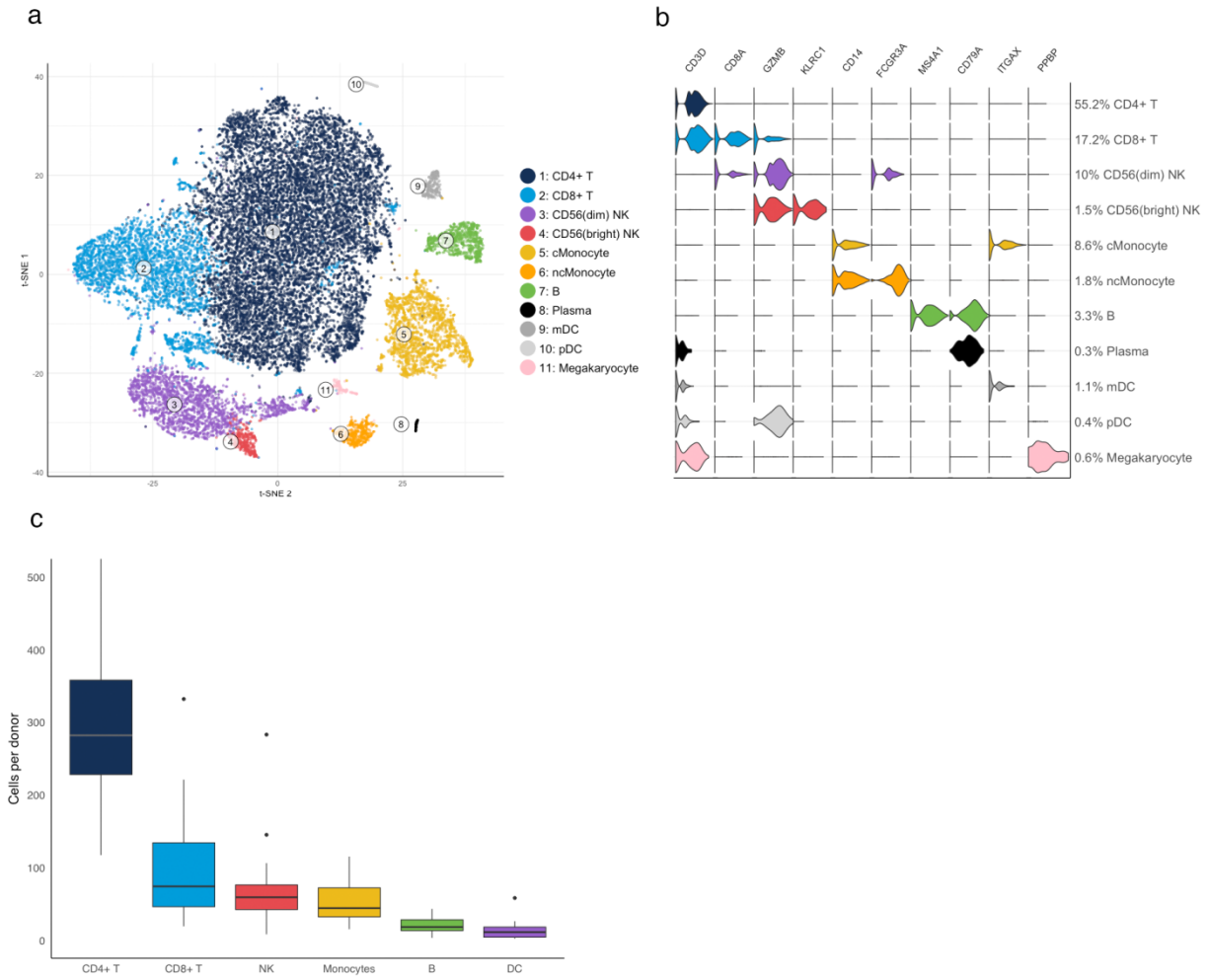
Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.

Supplementary Figure 1 van der Wijst et al.



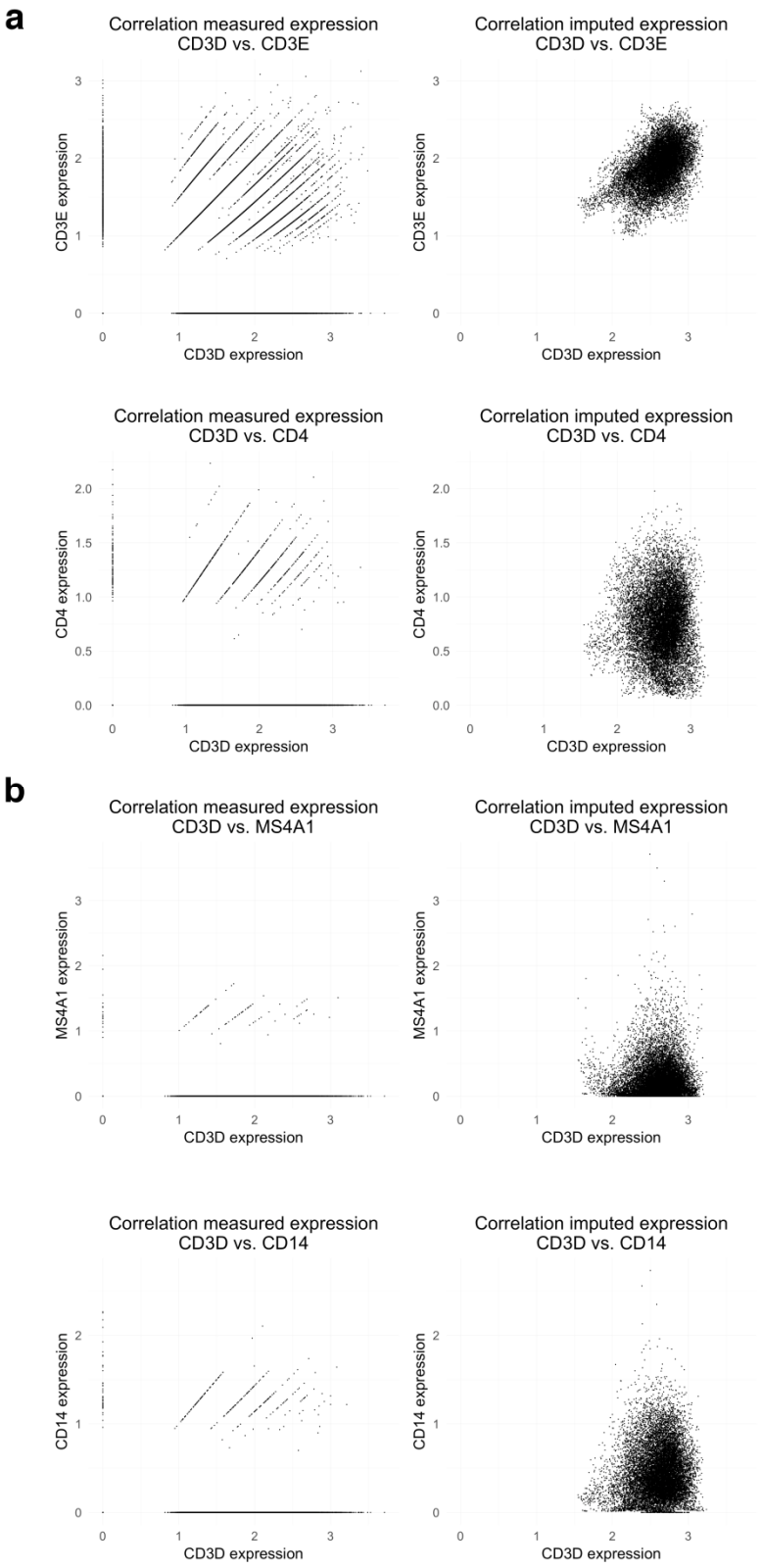
Supplementary Figure 1. Demultiplexing individuals, identification of doublets and cells failing QC. (**a**) PBMCs of six donors were pooled together in one sample pool. By taking into account the variable SNP positions between these donors, cells could be assigned to one (singlets) or to two donors (doublets). Results are shown for lane 1. (**b**) The number of reads mapping to the Y-chromosome (diamond: average) correlates with gender (M, male; F, female). The number in each boxplot indicates the number of identified singlets per donor after demultiplexing (results are shown for the six donors in lane 1 as in **a**). Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range. (**c**) Relation between the number of reads and (top) the fraction of genes mapping to the mitochondrial genome or (bottom) the number of genes. The percentage indicates the remaining cells (see **Suppl. Table 5**) that are removed by each cut-off. (**d**) Cells clustered in t-SNE space. Each dot represents a cell. Only the cells marked as singlet are used in downstream analysis, while all other colors indicate the reason for excluding these cells from further analysis. Only doublets of the lanes without sample mix-up are shown.

Supplementary Figure 2 van der Wijst et al.



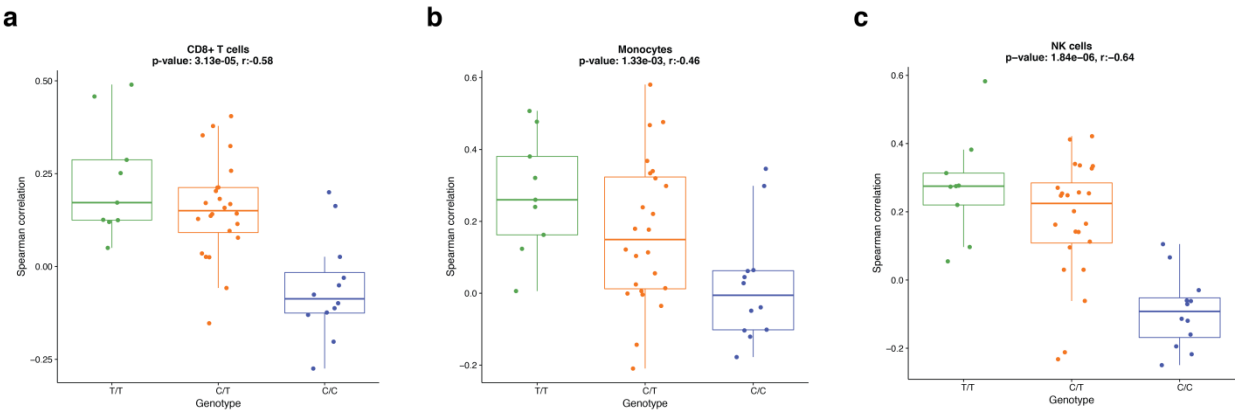
Supplementary Figure 2. Cell type classification and interindividual variability. (**a**) Cell clustering visualized in 2D space using t-SNE. Each dot represents a single cell. (**b**) Violin plots showing the expression of a selection of the used marker genes per cell type. (**c**) Interindividual variability in the number of cells detected per cell type. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range.

Supplementary Figure 3 van der Wijst et al.



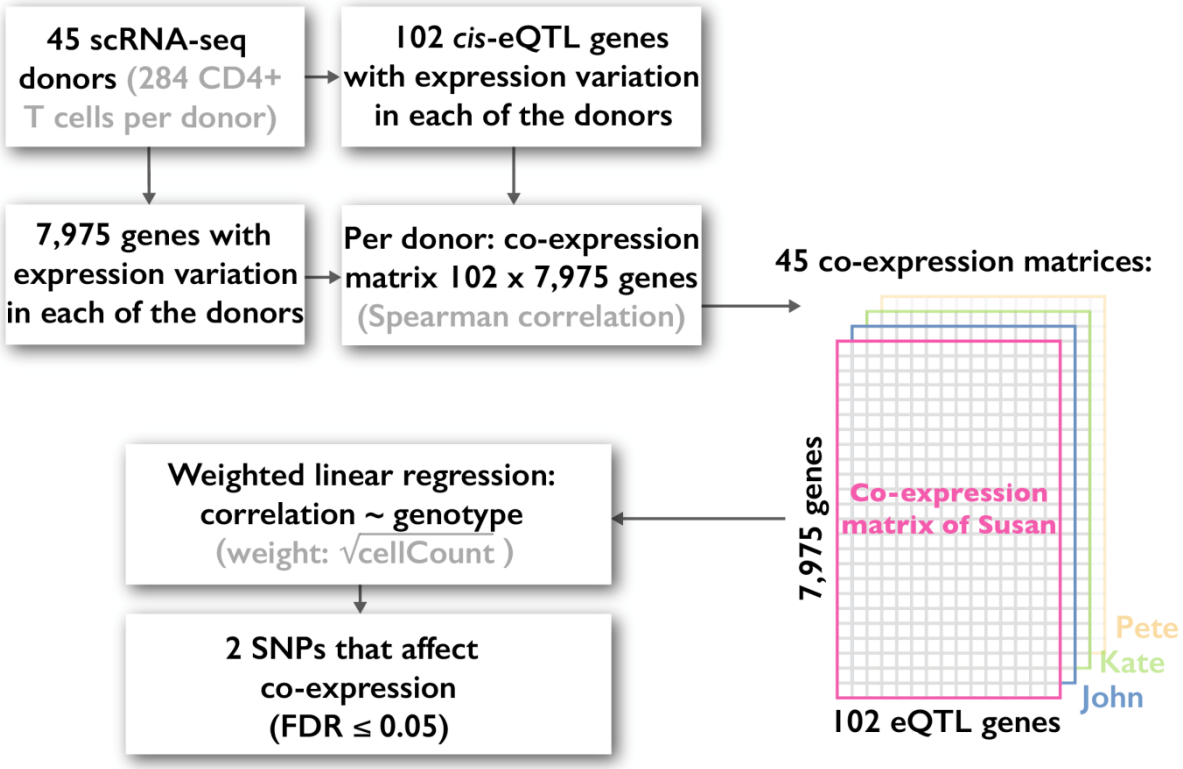
Supplementary Figure 3. Effect of MAGIC imputation on the gene expression in the CD4⁺ T cells. (**a**) Non-imputed and imputed expression of CD4⁺ T cell marker genes *CD3E* and *CD4* against the expression of *CD3D*. Expression for these genes is expected in all CD4⁺ T cells and after MAGIC imputation there are no more cells with zero expression. (**b**) Non-imputed and imputed expression of B cell marker *MS4A1* and monocyte marker *CD14* against the expression of *CD3D*. No expression of *MS4A1* and *CD14* is expected within CD4⁺ T cells, but after MAGIC imputation many cells show expression for these genes.

Supplementary Figure 4 van der Wijst et al.



Supplementary Figure 4. The most significant co-expression QTL effect (*RPS26-RPL21*) in other cell types than CD4+ T cells. The Spearman's rank correlation coefficient (r) between *RPS26* and *RPL21* expression stratified by SNP rs7297175 genotype per donor in (**a**) $CD8^+$ T cells, (**b**) monocytes and (**c**) NK cells. Each data point represents a single donor. Box plots show the median, the first and third quartiles, and 1.5 times the interquartile range.

Supplementary Figure 5 **van der Wijst et al.**



Supplementary Figure 5. Detailed workflow of the co-expression QTL analysis. A full description of the method can be found in the Online Methods section.

Supplementary Table 1. scRNA-seq eQTL analysis and concordance check confined to previously reported top-eQTLs from whole blood DeepSAGE and RNA-seq data

Supplied separately

All significant (at a gene-level FDR≤0.05) top SNP-gene combinations from whole blood RNA-seq (sheet1) or deepSAGE (sheet2) data that were detected in the bulk-like PBMC scRNA-seq sample. The “Concordant” column reveals whether these top SNP-gene combinations had either the same (“Yes”) or an opposite (“No”) allelic direction in the compared dataset. All columns correspond to the bulk-like PBMC scRNA-seq data, unless explicitly stated in the column name (e.g. “Z-score RNA-seq” or “Z-score deepSAGE”).

Supplementary Table 2. Genome-wide scRNA-seq eQTL analysis and replication of previously reported top-eQTLs from whole blood RNA-seq data

Supplied separately

Nominal p-values, significance (at a gene-level FDR≤0.05) and correlation of all significant (at a gene-level FDR≤0.05) top SNP-gene combinations identified in at least one of the following cell clusters within the scRNA-seq data: bulk-like PBMCs, CD4+ or CD8+ T-cells, NK-cells, B-cells, DCs, Monocytes or classical vs non-classical Monocyte subset. The “replication” column shows whether these top SNP-gene combinations had either the same (“Yes”) or an opposite (“No”) allelic direction, or were not significant/not tested (“Not found”) in the compared RNA-seq dataset.

Supplementary Table 3. Replication in purified cell type RNA-seq data of 19 eQTLs not found in the bulk-like PBMC scRNA-seq or whole blood RNA-seq data

Supplied separately

Nominal p-values, significance (at a gene-level FDR≤0.05) and correlation in each cell cluster for the 19 significant top SNP-gene combinations not identified in the bulk-like PBMC scRNA-seq sample or whole blood RNA-seq data. The four “replication” columns show whether these top SNP-gene combinations had either the same (“Yes”) or an opposite (“No”) allelic direction, or were not significant/not tested (“Not found”) in the compared RNA-seq dataset. For the replicated top SNP-gene combinations, the effect size and FDR are represented in separate columns.

Supplementary Table 4. Co-expression QTLs in the CD4$^+$ T cells.

Supplied separately

The most significant co-expression QTL for each of the 102 eQTL genes found in CD4+ T cells and with variance in expression in all 45 donors (sheet1). Of these 102, 3 eQTL genes are involved in 92 (P-value ≤ $1.27x10^{-7}$, corresponding to an eQTL-gene level FDR of 0.05) and 108 significant (P-value ≤ $4.72x10^{-7}$, corresponding to an eQTL-gene level FDR of 0.1) co-expression QTLs (sheet2). The columns provide the SNP id, eQTL gene Ensembl ID, eQTL gene HGNC name, interaction gene Ensembl ID, interaction gene HGNC name, assessed allele, nominal p-value of interaction model and the Spearman's rank correlation coefficient (r). In sheet1, the FDR is provided for the top co-expression QTLs. In sheet2, the nominal p-value and Spearman's rank correlation coefficient (r) are shown for the interaction in the MAGIC imputed gene expression data, whereas the nominal p-value and replication are given for the interaction in whole blood RNA-seq data. The Replication column shows whether co-expression QTLs were replicated ("Yes") or not ("No"), or could not be tested ("Not found"). In combination with the nominal p-value, one can extract whether non-replicated findings were due to non-significant results in the RNA-seq data or due to opposite allelic direction.

Supplementary Table 5. Sample pool information

Sample pool (SP)	Doublet		Singlet	Inconclusive	Total
	(#)	(%)			
1	191	4.4	4,097	21	4,309
2	673	17.5	3,166	5	3,844
3	1,051	21.0	3,941	6	4,998
4	97	2.9	3,285	10	3,392
5	116	4.0	2,739	13	2,868
6	130	3.5	3,552	3	3,685
7	84	2.6	3,107	7	3,198
8	52	2.0	2,505	4	2,561
Total	2,394		26,392	69	28,855

Supplementary Table 6. QC cut-offs

Metric	Loss	Remaining cells	
		(#)	(%)
Pre QC		28,855	100.0
Doublets	2,463	26,392	91.5
>5% mtDNA-encoded genes	1,075	25,317	87.7
>3500 genes	26	25,291	87.6

Supplementary Table 7. Cell type classification markers

Cell type	Subtype	(Relatively) High/present expression markers	(Relatively) Low/absent expression markers
CD4+ T cells		CD3D, CD3E, CD3G	CD8A, CD8B,
CD8+ T cells		CD3D, CD3E, CD3G, CD8A, CD8B, GZMB, PRF1	
NK cells	$CD56^{dim}$ $CD16^{+}$	FCGR3A, NKG7, GNLY, GZMB, PRF1	CD8A, CD8B
	$CD56^{bright}$ $CD16^{+/-}$	NKG7, GNLY, KLRC1	CD8A, CD8B, FCGR3A, GZMB, PRF1
Monocytes	$CD14^{bright}$ $CD16^{-}$ classical	CD14, LYZ, S100A9, CSF3R	FCGR3A, LYN, CSF1R, IFITM1, IFITM2, IFITM3 LYZ, S100A9, CSF3R
	$CD14^{dim}$ $CD16^{+}$ non-classical	CD14, FCGR3A, LYN, CSF1R, IFITM1, IFITM2, IFITM3	
B cells	Normal	CD79A	
	plasma	CD79A, MS4A1	
Dendritic cells	$CD1C^{+}$ myeloid	CD1C, ITGAX	CD14, CLEC4C
	Plasmacytoid	CLEC4C	CD14, CD1C, ITGAX
Megakaryocytes		GP9, ITGA2B, PF4, PPBP	

Supplementary Table 8. Sample metadata (including sample name, sample batch/lane of chip, gender and age)

Supplied separately

Supplementary Note 1 van der Wijst et al.

LifeLines Cohort Study – Author information

Ute Bultmann[1], J. M. (Marianne) Geleijnse[2], Pim van der Harst[3], Saakje Mulder[4], Judith G.M. Rosmalen[5], Elisabeth F.C. van Rossum[6], H.A. (Jet) Smit[7], Morris A Swertz[8,9], Evert A.L.M. Verhagen[10], Behrooz Z. Alizadeh[11], H.M. (Marike) Boezen[11], Lude Franke[8], Patrick Deelen[8,9], Gerjan Navis[12], Marianne G. Rots[13], Harold Snieder[11], Freerk van Dijk[8,9], Bruce H.R. Wolffenbuttel[14], Cisca Wijmenga[8].

1. University of Groningen, University Medical Center Groningen, Department of Social Medicine, Groningen, The Netherlands
2. Wageningen University, Department of Human Nutrition, Wageningen, The Netherlands
3. University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands
4. Lifelines Cohort Study, Groningen, The Netherlands
5. University of Groningen, University Medical Center Groningen, Interdisciplinary Center of Psychopathology of Emotion Regulation (ICPE), Department of Psychiatry, Groningen, The Netherlands
6. Erasmus Medical Center, Department of Endocrinology, Rotterdam, The Netherlands
7. University Medical Center Utrecht, Department of Public Health, Utrecht, The Netherlands
8. University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
9. University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands
10. VU Medical Center, Department of Public and Occupational Health, Amsterdam, The Netherlands
11. University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands
12. University of Groningen, University Medical Center Groningen, Department of Nephrology, Groningen, The Netherlands
13. University of Groningen, University Medical Center Groningen, Department of Medical Biology, Groningen, The Netherlands
14. University of Groningen, University Medical Center Groningen, Department of Endocrinology, Groningen, The Netherlands

Supplementary Note 2 van der Wijst et al.

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, ErasmusMC, Rotterdam, The Netherlands
4. Department of Genetic Epidemiology, ErasmusMC, 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, ErasmusMC, 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