SUPPLEMENTARY APPENDIX

Cell type-specific novel long non-coding RNA and circular RNA in the BLUEPRINT hematopoietic transcriptomes atlas

Luigi Grassi,[1,2,3*°] Osagie G. Izuogu,[4*] Natasha A.N. Jorge,[5] Denis Seyres,[1,2,3] Mariona Bustamante,[6,7,8] Frances Burden,[1,2,3] Samantha Farrow,[1,2,3] Neda Farahi,[9] Fergal J. Martin,[4] Adam Frankish,[4] Jonathan M. Mudge,[4] Myrto Kostadima,[1,2,4] Romina Petersen,[1,2] John J. Lambourne,[1,2] Sophia Rowlston,[1,2] Enca Martin-Rendon,[10,11] Laura Clarke,[4] Kate Downes,[1,2,3] Xavier Estivill,[12] Paul Flicek,[4] Joost H.A. Martens,[13] Marie-Laure Yaspo,[14] Hendrik G. Stunnenberg,[13] Willem H. Ouwehand,[1,2,3,15,16] Fabio Passetti,[5,17] Ernest Turro[1,2,3,18] and Mattia Frontini[1,2,16,19]

[1]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK; [2]National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge, UK; [3]National Institute for Health Research BioResource, Rare Diseases, Cambridge University Hospitals, Cambridge, UK; [4]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK; [5]Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil; [6]ISGlobal, Institute for Global Health, Barcelona, Spain; [7]Center for Genomic Regulation (CRG), Barcelona, Spain; [8]Universitat Pompeu Fabra, Barcelona, Spain; [9]Division of Respiratory Medicine, Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge, UK; [10]R&D Division, National Health Service (NHS)-Blood and Transplant, Oxford Centre, Oxford, UK; [11]Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, Oxford, UK; [12]Genes and Disease Research Group, Genetics and Genomics Program, Sidra Research Department, Sidra Medicine, Doha, Qatar; [13]Radboud University, Department of Molecular Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands; [14]Max Planck Institute for Molecular Genetics, Berlin, Germany; [15]Department of Human Genetics, the Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [16]British Heart Foundation Centre of Excellence, Cambridge Biomedical Campus, Cambridge, UK; [17]Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz, Curitiba, Brazil; [18]Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Cambridge, UK and [19]Institute of Biomedical & Clinical Science, College of Medicine and Health, University of Exeter Medical School, Exeter, UK

**LG and OGI contributed equally as co-first authors.*

°Current affiliation: Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK.

 doi:10.3324/haematol.2019.238147

Received: September 23, 2019.

Accepted: July 16, 2020.

Pre-published: July 23, 2020.

Correspondence: *MATTIA FRONTINI* - mf471@cam.ac.uk

ERNEST TURRO - et341@cam.ac.uk

Supplementary Information

Materials & Methods

Cell isolation

All samples were obtained from NHS Blood and Transplant blood donors and processed within 3 hours of collection or from cord blood donations at Rosie Hospital, Cambridge University Hospitals. Collections followed informed consent (ethical approval REC East of England 12/EE/0040). Detailed protocols, including antibody panels, are available at http://www.blueprint-epigenome.eu/. Briefly, neutrophils and monocytes were isolated from peripheral blood whole units (460 ml) or from cord blood units. Peripheral blood mononuclear cells (PBMCs) were separated by gradient centrifugation (Percoll 1.078 g/ml) whilst neutrophils were isolated from the pellet, after red blood cell lysis, by CD16 positive selection (Miltenyi). PBMCs were further separated using a second gradient (Percoll 1.066 g/ml) to obtain a monocyte rich layer. Monocytes were further purified by CD16 depletion followed by CD14 positive selection (Miltenyi).

The purification of macrophages (M0), LPS activated macrophages (M1), alternatively activated macrophages (M2), endothelial cell precursors, erythroblasts, megakaryocyte, naive B lymphocytes, naive CD4 lymphocytes and naive CD8 lymphocytes used in this study has been described extensively[1-4]. Regulatory CD4 lymphocytes (T regs), CD4 central memory lymphocytes (CM) and CD4 effector memory lymphocytes (EM) were isolated by flow activated cytometry (FACS) using the following surface markers combinations: T regs, CD3+ CD4+ CD25+ CD127low; CD4 CM, CD3+ CD4+ CD45RA- CD62L+; CD4 EM, CD3+ CD4+ CD45RA- CD62L-. CD8 central memory lymphocytes (CM), CD8 EM and CD8 terminally differentiated effector memory lymphocytes (TDEM) were isolated by FACS using the following surface markers combinations: CD8 CM, CD3+ CD8+ CD62L+ CD45RA-; CD8 EM, CD3+ CD8+ CD62L- CD45RA-; CD8 TDEM, CD3+ CD8+ CD62L- CD45RA+. B memory lymphocytes and B class switch lymphocytes were isolated by FACS, using the following surface markers combinations: B memory, CD19+ CD27+ IgD+; B class switch, CD19+ CD27+ IgD- CD38dim. Natural Killer cells (NK) were isolated by FACS using the following surface markers: CD3- CD56dim CD16+. Eosinophils and basophils were isolated from a mixed leukocytes pellet obtained by sedimentation of whole blood 6% hydroxyethyl starch (Grifols, Cambridge, UK) for 30 minutes using EasySep (Stemcell Technologies) as previously described[5]. Monocyte derived dendritic cells were generated from cord blood CD34 depleted PBMCs after a second Percoll gradient (1.066 g/ml) to enrich monocytes using a PromoCell dendritic cell isolation kit. Bone marrow derived mesenchymal stem cell isolation has been previously described[6]. Platelets were isolated from platelet rich

plasma after leukocyte (CD45+) depletion as previously described[2]. The purity of each cell fraction was assessed by flow cytometry and/or morphological analysis after cytospin preparation and staining. The purified cells were resuspended in Trizol. Samples that did not meet predefined criteria of cell purity (>95%) were not sent for sequencing.

RNA extraction

RNA was extracted from TRIzol according to the manufacturer's instructions, quantified using a Qubit RNA HS kit (Thermofisher) and quality controlled using a Bioanalyzer RNA pico kit (Agilent).

Library construction and sequencing

With the exception of platelets, eosinophils and basophils, libraries were prepared using a TruSeq Stranded Total RNA Kit with Ribo-Zero Gold (Illumina) using 200ng of RNA as input. Platelet, eosinophil and basophils samples were prepared with the Kapa stranded RNA-seq kit with riboerase (Roche) according to the manufacturer's instructions.

Small RNA extraction

RNA was extracted using the miRNeasy Mini Kit (Qiagen) from cell pellets provided their RNA Integrity Number (RIN) was between 7.3 and 10, as assessed with an RNA 6000 Nano kit on a 2100 Bioanalyzer (Agilent). Small RNA libraries were prepared using the NEBNext® Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs) and the LongAmp Taq 2x Master Mix. Size selection was performed with 6% polyacrylamide gels, and library quality was verified on a 2100 Bioanalyzer (Agilent). Equimolar (2 nM) amounts of each library, as verified with Picogreen® dsDNA Quantification Reagent (Promega), were pooled and sequenced on an Illumina HiSeq 2000 using 50 bp single end reads.

CircRNA identification and comparisons

Backsplice junctions were identified using CIRI[7], CIRCexplorer[8], find_circ[9], circRNA_finderP[10] and PTESFinder[11] (parameters: JSpan=10, PID=0.85, segment_size=65), mapping against the human genome build GRCh37. Junctions called by fewer than three methods were removed. The genomic positions of backsplice junctions were compared to previously identified junctions in circbase[12] (obtained 05/2018), annotated splice sites in Ensembl 75 and known segmental duplications[13] in the genome. Backsplice junctions overlapping multiple genes, readthrough transcripts or segmental duplications were excluded from downstream analyses.

Backsplice classification

Backsplices were classified into five groups based on their genomic locations relative to

Ensembl 75 annotations *exonic_known:* the backsplice corresponds to known splice sites; *exonic_novel:* the backsplice overlaps at least one annotated exon and utilises only one known splice site; *intronic:* the backsplice is internal to an annotated intron; *intergenic:* the backsplice does not overlap any annotated exons or introns, and *antisense:* the backsplice is antisense to annotated exons or introns.

Modelling of circRNA expression.

The read counts reported by PTESFinder were normalised by dividing them by the total number of splicing reads in each sample and multiplied by 10^6. For each sample, we computed the abundance proportion (AP) of a gene as the number of backsplice reads in that gene divided by the total number of spliced reads of any kind in the sample. Differential expression analysis was performed using DESeq2[14]. Z-scores for differentially expressed backsplices identified by DESeq2 were computed over samples from the normalised backsplice read counts.

Supplementary Figures.

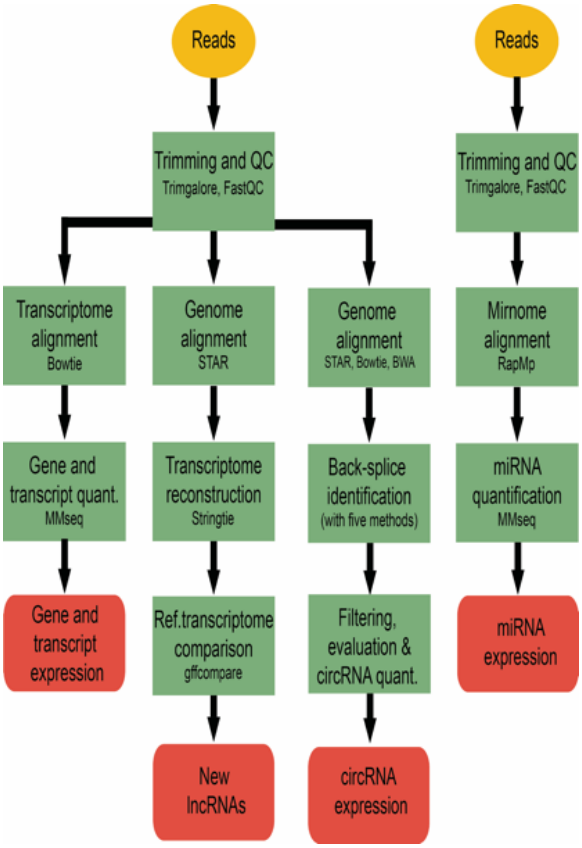


Figure S1: Schematic of the bioinformatic pipelines. Paired end reads (75bp and 150bp; left), were trimmed to remove adapter sequences, quality controlled and aligned to the human transcriptome with Bowtie. The alignments were modelled by MMSEQ to estimate gene and transcript level expression. The reads were also aligned to the human genome with STAR and fed to StringTie to identify novel genes. The reads were aligned to the human genome with three aligners (STAR, Bowtie and BWA) to identify circRNA species and model their expression levels. Single end reads (50bp; right) were aligned to the human mirnome with RapMap. The alignments were modelled by MMSEQ to estimate miRNA expression levels.

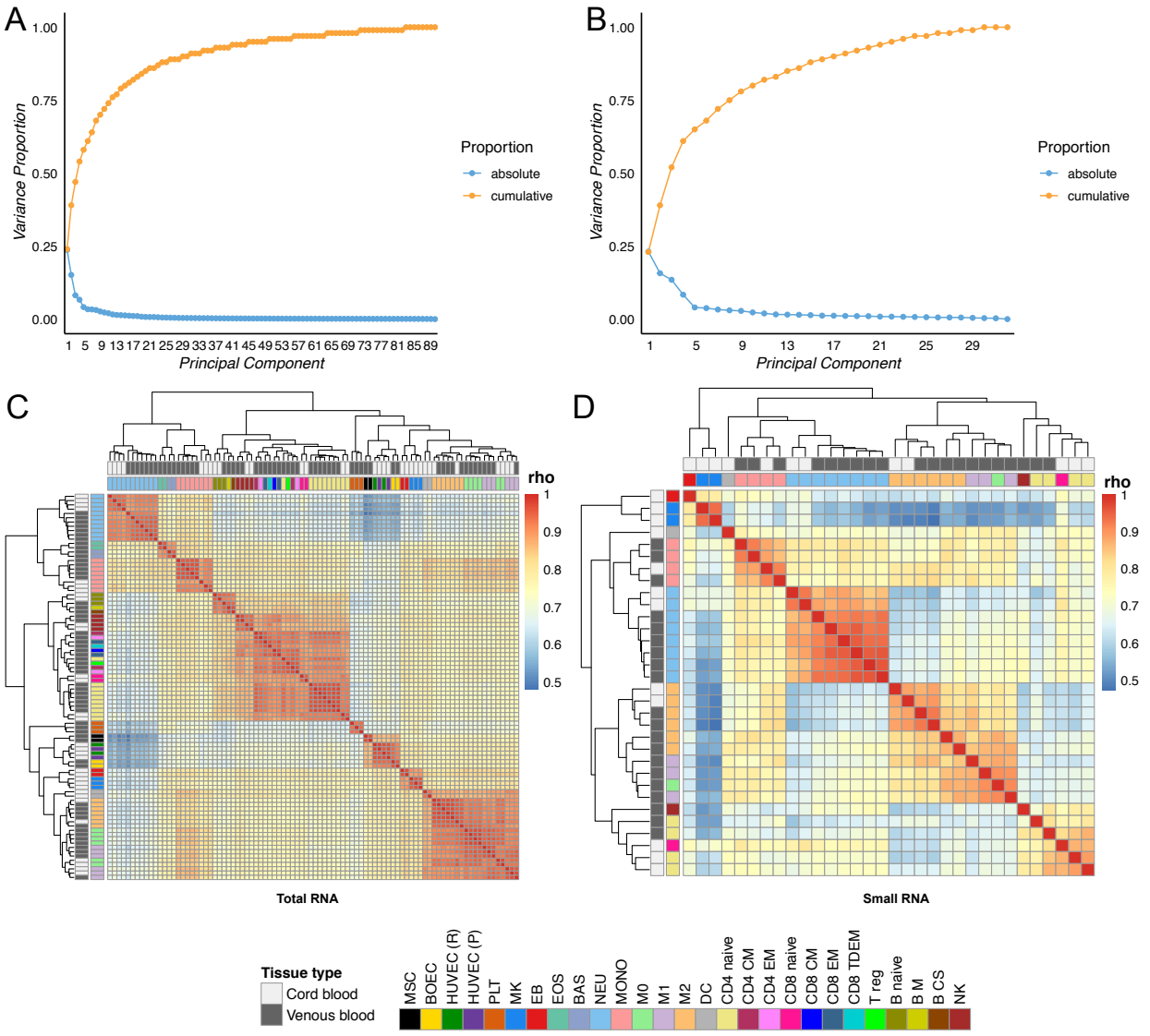


Figure S2: Principal components analysis and hierarchical clustering of gene and miRNA expression. S2A: Proportion and cumulative proportion of variance explained by successive principal components, derived from the log expression estimates of genes with a log expression estimate greater than zero in at least one sample. **S2B**: Proportion and cumulative proportion of variance explained by successive principal components, derived from the log expression estimates of miRNAs with a unique read count >10 in at least one sample. **S2C**: Heatmap of the Spearman rank correlation coefficient (rho) of genes with a log expression estimate greater than zero in at least one sample. The rows and columns have been ordered by complete linkage hierarchical clustering using 1-rho as the distance measure. **S2D**: Heatmap of the Spearman rank correlation coefficient (rho) of miRNAs with unique read count >10 in at least one sample. The rows and columns have been ordered by complete linkage hierarchical clustering using 1-rho as the distance measure.

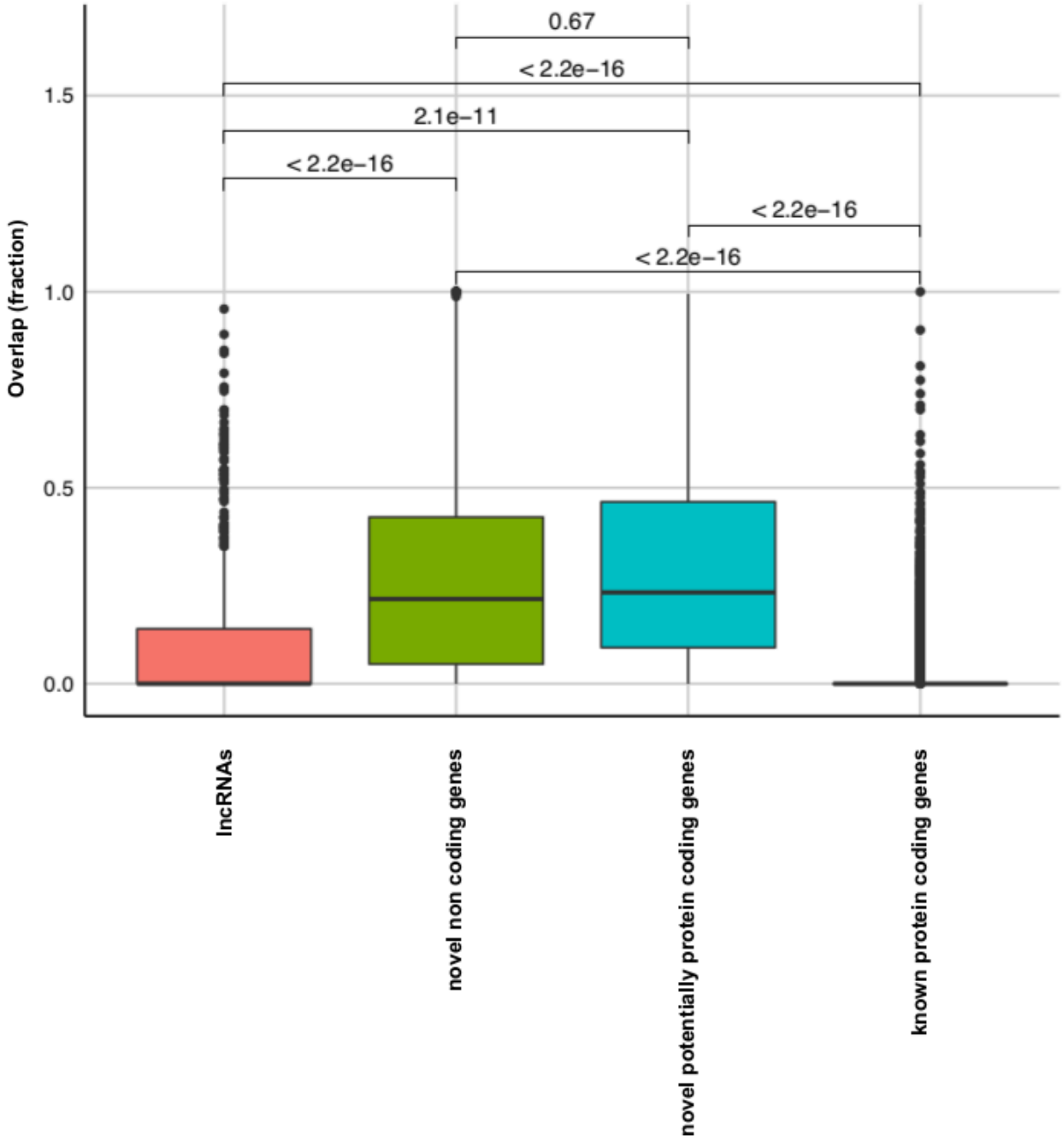


Figure S3: Novel non-coding and potentially protein coding genes overlap with transposon-associated regions and other repetitive or low complexity regions. Box plots of the fraction of the genomic annotated lncRNAs, novel non-coding genes, novel potentially protein coding genes and known protein coding genes which overlap transposon-associated regions and other repetitive or low complexity regions. The centre mark and lower and upper hinges of the boxplots respectively indicate the median, 25th and 75th percentiles. Outliers beyond 1.5 times the interquartile range from each hinge are shown. Pairwise Wilcoxon signed-rank test *P* values Bonferroni corrected are reported.

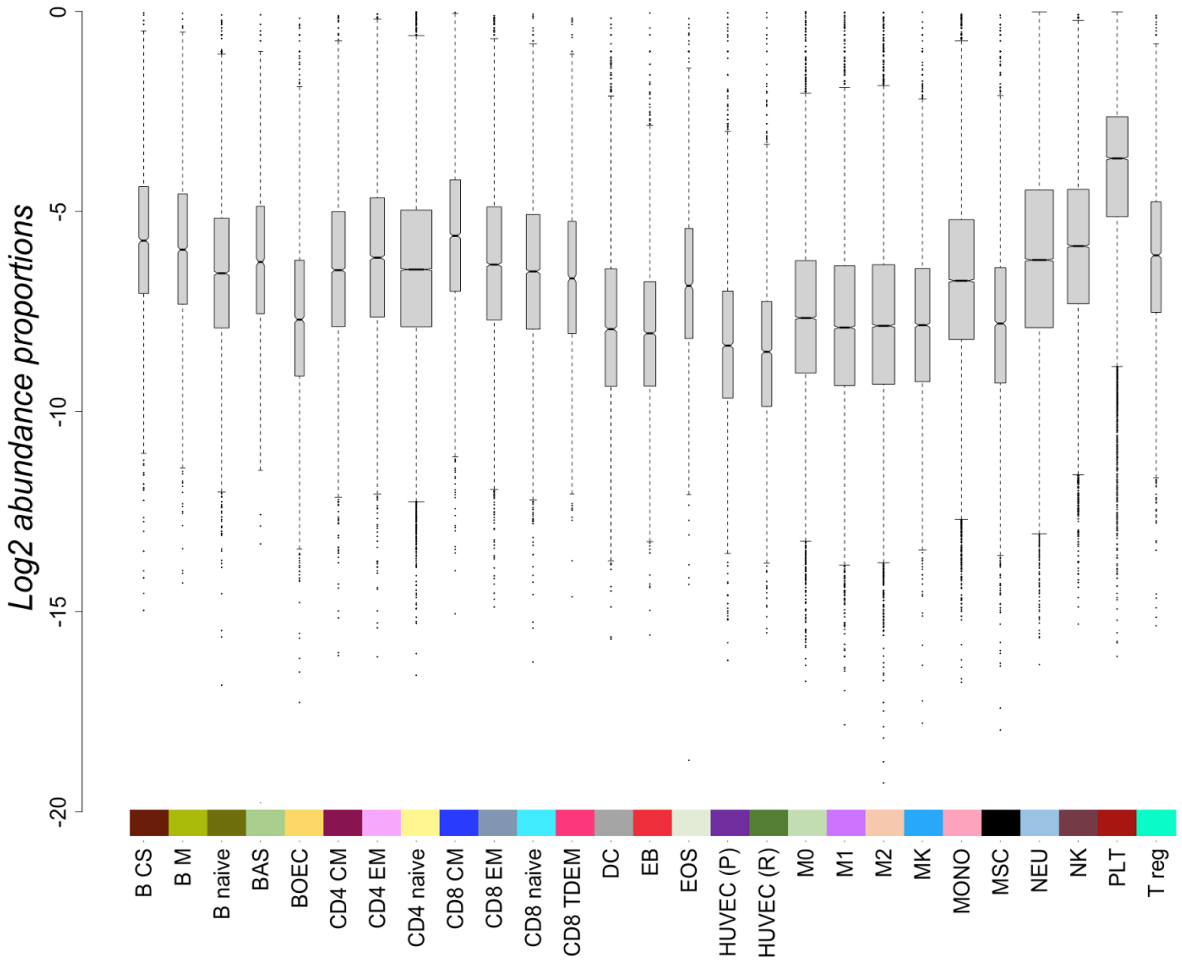


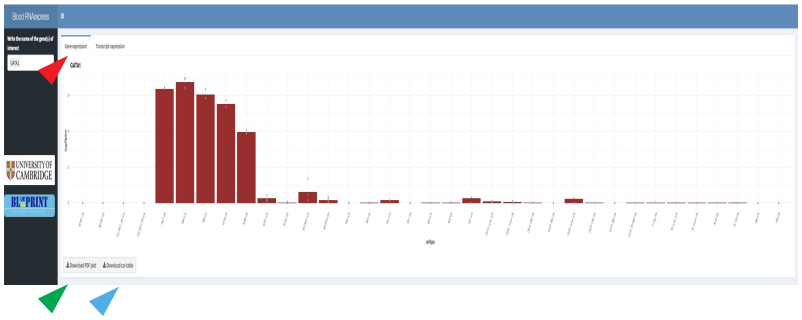
Figure S4: circRNA abundance in blood cells. Box plots of circRNA abundance proportions in each type of blood cell. Abundance proportions were derived by dividing total backsplice read counts with total splice reads from host genes. The width of each box is proportional to the number of samples of each cell type. The centre mark and lower and upper hinges of the boxplots respectively indicate the median, 25th and 75th percentiles. Outliers beyond 1.5 times the interquartile range from each hinge are shown.

A

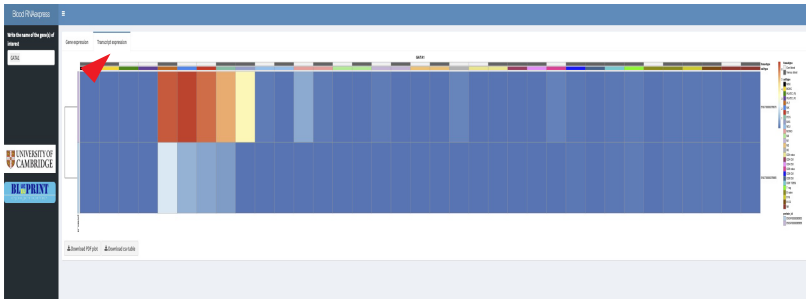

Blood Atlas
Genes/Transcripts
miRNAs
Novel Genes/Transcripts
circRNAs

B

C



D



E

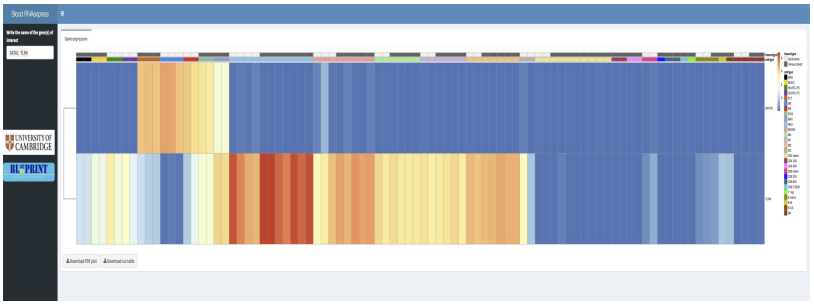


Figure S5: Overview of the web application's functionality. S5A: homepage, the user selects one of four options, red arrow. **S5B**: independently of the option chosen, the user enters the name of the gene for which he wishes to visualise expression (red arrow). S5C: visualisation of the expression levels of a single gene (switch to transcript expression by selecting the tab indicated by the red arrow). Each bar indicates the median $\log_2$ expression+1 in a cell type. The dots show the individual values for the available replicates. The graphical representation and the expression values can be downloaded as a PDF (green arrow) and a csv file (blue arrow), respectively. GATA1 is used as an example. **S5D:** visualisation of transcript expression for a single gene as a heatmap, where each row represents a different transcript originating from the queried gene. GATA1 is used as an example **S5E:** visualisation of gene expression for multiple genes as a heatmap showing median expression in each cell type. Each row represents one of the queried genes. GATA1 and TLR4 are depicted as examples.

List Supplementary Tables

S1 List of total RNA-sequencing samples used in this manuscript.

S2 List of abbreviations and grouping used in this manuscript.

S3 List of small RNA-sequencing samples used in this manuscript.

S4 Number of genes accounting for 50% and 75% of total expression.

S5 List of GO terms found enriched for the genes in S4.

S6 List of gene constituting the transcriptional signature of each cell type.

S7 List of GO terms found enriched for the genes in S6.

S8 List of miRNAs constituting the transcriptional signature of each cell type.

S9 List of circRNA back splice junctions identified.

S10 Classification of circRNA back splice junctions identified.

S11 List circRNA abundance rations.

Supplementary files description

Supplementary file 1 lists the protein coding genes accounting for 50% and 75% of the transcriptional output of each cell type.

Supplementary file 2 contains the miRNAs accounting for 75% of the small RNA transcriptome.

Supplementary file 3 contains the genomic coordinates of the novel genes identified by guided transcriptome reconstruction.

Supplementary file 4 contains the differentially expressed circRNAs found by pairwise comparisons of functional categories of cell type.

References

1. Chen L, Kostadima M, Martens JHA, et al. Transcriptional diversity during lineage commitment of human blood progenitors. Science. 2014 Sep 26;345(6204):1251033. doi:10.1126/science.1251033. Cited in: Pubmed; PMID 25258084.

2. Astle WJ, Elding H, Jiang T, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell. 2016 Nov 17;167(5):1415-1429 e1419. Epub 2016/11/20. doi:10.1016/j.cell.2016.10.042. Cited in: Pubmed; PMID 27863252.

3. Petersen R, Lambourne JJ, Javierre BM, et al. Platelet function is modified by common sequence variation in megakaryocyte super enhancers. Nat Commun. 2017 Jul 13;8:16058. doi:10.1038/ncomms16058. Cited in: Pubmed; PMID 28703137.

4. Farlik M, Halbritter F, Muller F, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell. 2016 Dec 1;19(6):808-822. doi:10.1016/j.stem.2016.10.019. Cited in: Pubmed; PMID 27867036.

5. Brode S, Farahi N, Cowburn AS, et al. Interleukin-5 inhibits glucocorticoid-mediated apoptosis in human eosinophils. Thorax. 2010 Dec;65(12):1116-1117. doi:10.1136/thx.2009.124909. Cited in: Pubmed; PMID 20805156.

6. Martin-Rendon E, Sweeney D, Lu F, et al. 5-Azacytidine-treated human mesenchymal stem/progenitor cells derived from umbilical cord, cord blood and bone marrow do not generate cardiomyocytes in vitro at high frequencies. Vox Sang. 2008 Aug;95(2):137-148. doi:10.1111/j.1423-0410.2008.01076.x. Cited in: Pubmed; PMID 18557828.

7. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. 2015 Jan 13;16:4. doi:10.1186/s13059-014-0571-3. Cited in: Pubmed; PMID 25583365.

8. Dong R, Ma XK, Chen LL, Yang L. Genome-Wide Annotation of circRNAs and Their Alternative Back-Splicing/Splicing with CIRCexplorer Pipeline. Methods Mol Biol. 2019;1870:137-149. Epub 2018/12/13. doi:10.1007/978-1-4939-8808-2_10. Cited in: Pubmed; PMID 30539552.

9. Hansen TB, Veno MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res. 2016 Apr 7;44(6):e58. Epub 2015/12/15. doi:10.1093/nar/gkv1458. Cited in: Pubmed; PMID 26657634.

10. Chen L, Yu Y, Zhang X, et al. PcircRNA_finder: a software for circRNA prediction in plants. Bioinformatics. 2016 Nov 15;32(22):3528-3529. Epub 2016/08/06. doi:10.1093/bioinformatics/btw496. Cited in: Pubmed; PMID 27493192.

11. Izuogu OG, Alhasan AA, Alafghani HM, et al. PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events. BMC Bioinformatics. 2016 Jan 13;17:31. doi:10.1186/s12859-016-0881-4. Cited in: Pubmed; PMID 26758031.

12. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014 Nov;20(11):1666-1670. doi:10.1261/rna.043687.113. Cited in: Pubmed; PMID 25234927.

13. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001 Jun;11(6):1005-1017. doi:10.1101/gr.gr-1871r. Cited in: Pubmed; PMID 11381028.

14. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8. Cited in: Pubmed; PMID 25516281.