ARTICLE

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.

Ferrata Storti Foundation

Haematologica 2021
Volume 106(10):2613-2623

ABSTRACT

Transcriptional profiling of hematopoietic cell subpopulations has helped to characterize the developmental stages of the hematopoietic system and the molecular bases of malignant and non-malignant blood diseases. Previously, only the genes targeted by expression microarrays could be profiled genome-wide. High-throughput RNA sequencing, however, encompasses a broader repertoire of RNA molecules, without restriction to previously annotated genes. We analyzed the BLUEPRINT consortium RNA-sequencing data for mature hematopoietic cell types. The data comprised 90 total RNA-sequencing samples, each composed of one of 27 cell types, and 32 small RNA-sequencing samples, each composed of one of 11 cell types. We estimated gene and isoform expression levels for each cell type using existing annotations from Ensembl. We then used guided transcriptome assembly to discover unannotated transcripts. We identified hundreds of novel non-coding RNA genes and showed that the majority have cell type-dependent expression. We also characterized the expression of circular RNA and found that these are also cell type-specific. These analyses refine the active transcriptional landscape of mature hematopoietic cells, highlight abundant genes and transcriptional isoforms for each blood cell type, and provide a valuable resource for researchers of hematologic development and diseases. Finally, we made the data accessible via a web-based interface: *https://blueprint.haem.cam.ac.uk/bloodatlas/*.

Correspondence:

MATTIA FRONTINI
mf471@cam.ac.uk

ERNEST TURRO
et341@cam.ac.uk

Received: September 23, 2019.

Accepted: July 16, 2020.

Pre-published: July 23, 2020.

https://doi.org/10.3324/haematol.2019.238147

Introduction

Knowledge of the transcriptional programs underpinning the diverse functions of hematopoietic cells is essential for understanding how and when these functions are performed and for resolving the molecular bases of hematologic diseases. Thanks to its accessibility, blood is the tissue of choice for the implementation of novel assays in primary samples. Indeed, several studies aiming to characterize gene expression profiles in the post-genome era have been performed on increasingly purified primary hematopoietic cell populations.[1-3] These studies used expression arrays and thus required prior specification of the sequences to be interrogated. The probed sequences were often derived from the analysis of a very limited number of tissues and cell types,[4] despite the early discovery that transcription is widespread throughout the human genome.[5] The introduction of high-throughput nucleic acid sequencing technologies[6] has improved the assembly of the human genome and the annotation of transcriptomes therein, and has enabled a more comprehensive analysis of gene expression using transcriptomic assembly approaches.[7] The BLUEPRINT consortium[8] was established to characterize the epigenetic state and transcriptional profile of different types of hematopoietic cells. Reference datasets for DNA methylation, histone modifications and gene expression were generated from highly purified cell populations using state-of-the-art technologies, in accordance with quality standards set by the International Human Epigenome Consortium.[9] RNA-sequencing data from over 270 samples encompassing 55 cell types have been made publicly available (*http://dcc.blueprint-epigenome.eu*). A subset of these data has been described previously.[10,11] Here, we present the analysis of 90 total RNA samples obtained from cord and adult peripheral blood, each consisting of one of 27 mature cell types and 32 small RNA samples, each consisting of one of 11 mature cell types. We used a Bayesian differential expression analysis approach[12,13] to determine changes in the expression levels of genes and transcripts at lineage commitment stages and to identify cell type-specific transcriptional signatures. We performed guided transcriptome reconstruction[7] using total RNA-sequencing reads, identifying 645 multi-exonic transcripts originating from 400 intergenic novel genes. The majority of the novel transcripts had low protein coding potential and high cell type specificity. Additionally, we identified 55,187 circular RNA (circRNA), which also displayed high cell type specificity, highlighting the potential role of non-coding transcripts in hematopoiesis. To enable exploration and reuse of the data by the biomedical community, we developed a web interface for plotting expression patterns of genes and transcripts and downloading normalized expression data (*https://blueprint.haem.cam.ac.uk/bloodatlas/*).

Methods

Ethical approval

Samples were obtained from National Health Service Blood and Transplant blood donors and from cord blood donations at Cambridge University Hospitals, following informed consent. Ethical approval was obtained for A Blueprint of Blood Cells (REC East of England 12/EE/0040).

Cell isolation, RNA extraction and library construction

The protocols used for cell isolation, RNA extraction and library construction are described in the *Online Supplementary Material*.

Bioinformatic analysis

An overview of the bioinformatic pipeline is shown in *Online Supplementary Figure S1*. To analyze the expression of known genes and transcripts, we trimmed reads with Trim Galore (v0.3.7; parameters "-q 15 -s 3 --length 30 -e 0.05") and aligned them to Ensembl v75 of the human transcriptome with Bowtie[14] (1.0.1; parameters "-a --best --strata -S -m 100 -X 500 --chunkmbs 256 --nofw --fr"). Small RNA-sequencing reads were also trimmed with Trim Galore (v0.3.7; parameters "-f fastq -e 0.05 -q 15 -O 3") and aligned to the miRBase (v21) human mature microRNA (miRNA) with RapMap (v 0.4.0) using the parameters "quasimap -c -s -z 0.9". We used MMSEQ[12] and MMDIFF[13] (v1.0.10; default parameters) to estimate gene, transcript and miRNA expression levels, and to identify features that were differentially expressed across cell types. This choice of methodology allowed us to obtain regularized transcript and gene-level posterior estimates of expression and the corresponding measures of posterior uncertainty, which could then be accounted for in the modeling of differential expression. For guided transcriptome assembly, we used STAR (v2.4.1c; parameters "--runThreadN 8 --outStd SAM --outSAMtype BAM Unsorted --outSAMstrandField intronMotif") to align trimmed reads to build GRCh37 of the reference human genome. We sorted the bam files by coordinate and indexed them with samtools (v 1.3.1).[15] We performed guided transcriptome assembly for each sample using StringTie7 (v 1.3.4; parameters "-p 8 --rf -G Ensembl_75.gtf -v -l BPSTRG"). We also used StringTie to combine these transcriptomes into a single merged transcriptome, which we then compared to the annotations in Ensembl 75 using Gffcompare.[16] We identified intergenic transcripts and filtered out the ones overlapping known transcripts annotated in Gencode (v19)[17] and UCSC (v hg19)[18] using the GenomicRanges package.[19] We assessed the protein coding potential of the novel intergenic multi-exonic transcripts using the Coding-Potential Assessment Tool (CPAT) (v 1.2.4).[20] We chose CPAT because of its superior accuracy relative to competing methods.[20] A coding potential >0.364 was considered to discriminate between protein-coding and non-coding transcripts, in accordance with the human-specific guidance in the CPAT manual (*http://rna-cpat.sourceforge.net/*). We estimated the expression levels of novel genes and transcripts using MMSEQ, as described above for known genes and transcripts. We computed the expression specificity parameter Tau[21] to compare the cell type specificities of novel genes, known long non-coding RNA (lncRNA) and known protein-coding genes. We used the BioConductor R package "phastCons100way. UCSC.hg19"[22] to obtain sequence conservation scores of novel genes, known lncRNA and known protein-coding genes. A detailed description of the computational methods used to identify circRNA, compare their sequences to known sequences and quantify expression levels is given in the *Online Supplementary Material*.

Data availability

All data used in this manuscript are available from the European Genome-phenome Archive (EGA) (*https://www.ebi.ac.uk/ega/dacs/EGAC00001000135I*). The dataset identities are listed in *Online Supplementary Table S1*. Links to the datasets at EGA are also available from the BLUEPRINT data access portal (*http://dcc.blueprint-epigenome.eu/#/datasets*).

Results

Transcriptome complexity of hematopoietic cell types

We isolated 90 samples (Figure 1A and B, *Online Supplementary Table S1*) from 72 whole blood and cord blood donations, either by magnetic bead separation or by flow activated cell sorting (*Online Supplementary Methods*). Total RNA data were generated from the following 27 cell types: erythroblasts (EB), megakaryocytes (MK), platelets (PLT), eosinophils (EOS), basophils (BAS), neutrophils (NEU), monocytes (MONO), non-activated macrophages (M0), lipopolysaccharide activated macrophages (M1), alternatively activated macrophages (M2), dendritic cells (DC), naive CD4 lymphocytes (CD4 naive), central memory CD4 lymphocytes (CD4 CM), effector memory CD4 lymphocytes (CD4 EM), regulatory CD4 lymphocytes (TREG), naive CD8 lymphocytes (CD8 naive), central memory CD8 lymphocytes (CD8 CM), effector memory CD8 lymphocytes (CD8 EM), terminally differentiated effector memory CD8 lymphocytes (CD8 TDEM), naive B lymphocytes (B naive) , memory B lymphocytes (B M), class switch B lymphocytes (BCS), natural killer cells (NK), blood outgrowth endothelial cell progenitors (BOEC), umbilical vein endothelial cells (resting and proliferating; HUVEC R and P) and mesenchymal stem cells (MSC). Small RNA data were generated from the following 11 cell types: EB, MK, NEU, MONO, M0, M1, M2, DC, CD4 naive, CD8 naive and NK. An overview of the number of samples assayed of each cell type by total and small RNA-sequencing is presented in Figure 1A and B and *Online Supplementary Table S2*. We generated a mean of 91M 75 bp paired-end reads for total ribosomal RNA-depleted samples, except for platelets (PLT), basophils (BAS) and eosinophils (EOS), which were sequenced at a comparable depth but with 150 bp paired-end reads (*Online Supplementary Table S1*). We also generated a mean of 4.5M 50 bp single-end reads for small RNA samples (*Online Supplementary Table S3*). Principal component analysis of the log expression estimates for both protein-coding genes and small RNA showed distinct clustering by cell type according to their ontology along the first two principal components, which explained approximately 40% of the variance in expression of both types of RNA species (Figure 1C and D, *Online Supplementary Figure S2A* and *B*). This correspondence was also apparent by hierarchical clustering of samples using Spearman rank correlation (*Online Supplementary Figure S2C* and *D*).

The GTEx project[23] showed that whole blood has a very low gene expression complexity compared to that of other tissues, as 60% of all blood transcripts emanate from three hemoglobin genes.[24] However, a low complexity of a heterogeneous tissue may mask a high complexity of some of its component cell types. We therefore analyzed transcriptome complexity in different types of blood cells. After excluding mitochondrial genes from the analysis to account for their considerable variation in steady-state expression across individuals,[25] the number of protein-coding genes contributing 50% of total expression ranged from only 14 in PLT to 600 in BAS. The number of protein-coding genes contributing 75% of total expression ranged from 168 in PLT to 2,422 in resting HUVEC (Figure 2A, *Online Supplementary Table S4, Online Supplementary File 1*). With the exception of PLT, the sets of genes yielding 75% of total expression in each cell type showed enrichment for gene ontology (GO) terms only for functional categories related to general biological processes, such as translation or transcription. Thus, cellular integrity and basic cellular functions are supported at the transcriptional level even in mature cell types, some of which have short half-lives. In PLT, however, we found an enrichment for GO terms related to the core functions of platelets (i.e., hemostasis, wound healing, coagulation, platelet degranulation), while more general processes featured less prominently (*Online Supplementary Table S5*). The corresponding analysis of the small RNA data showed a very low complexity: between one and seven miRNA accounted for 50% of total expression and fewer than ten miRNA accounted for 75% of the expression in each of the 11 cell types (Figure 2B, *Online Supplementary File 2*).

Transcriptional signatures correspond to hematopoietic cell functions

As the most highly transcribed genes in a given cell type are in general not enriched for GO terms describing that cell type's specific functions, we reasoned that these functions must be encoded primarily by other more lowly expressed genes. The expression levels of these genes should in principle correlate with cell type in order to ensure function specialization. To determine which genes form the transcriptional signature of each cell type, we grouped cell types into functional categories (*Online Supplementary Table S2*) and then identified heterogeneously expressed genes over these categories through a Bayesian comparison of two statistical models: one in which the gene under consideration had a global mean expression parameter and another in which the gene had a different mean expression parameter for each category. Both models included a binary covariate accounting for the source of the blood samples (venous or cord). Using this approach, we found that 19,861 (59.5%) of HUGO Gene Nomenclature Committee (HGNC)-annotated genes had a posterior probability of differential expression >0.8. Over half of these differentially expressed genes had a mean log expression across samples >0. In contrast, only 3.5% of the non-differentially expressed genes had a mean log expression >0, indicating that the number of ubiquitously expressed housekeeping genes in hematopoiesis is a few hundred. The differentially expressed genes were then classified by the cell type in which their expression was greatest. To ensure that the classification recapitulated cellular functions specific to the mature blood cells in this atlas, rather than functions of shared progenitors from which they originate, we only classified the 16,572 genes whose maximum log expression level was at least 0.1 (i.e., 10.5%) greater than that found in the cell type with the second greatest expression (*Online Supplementary Methods, Online Supplementary Table S6*). For example, VWF was assigned the endothelial cell (EC) label because, firstly, its expression varies across cell types (posterior probability of differential expression approximately = 1), secondly, VWF is most highly expressed in EC (log expression estimate = 6.0) and, thirdly, the second highest expressed category (MK/PLT, combined because PLT are the immediate anucleated descendants of MK) has a log expression estimate (averaged over MK and PLT) of 2.2, which is smaller than 6.0 by more than 0.1 units (Figure 3A). The number of genes assigned to each category ranged from 186 in CD8 T lymphocytes (CD8TC) to 3,502 in MK/PLT (Figure 3B). Using these groups of genes, we found enrichment for GO terms

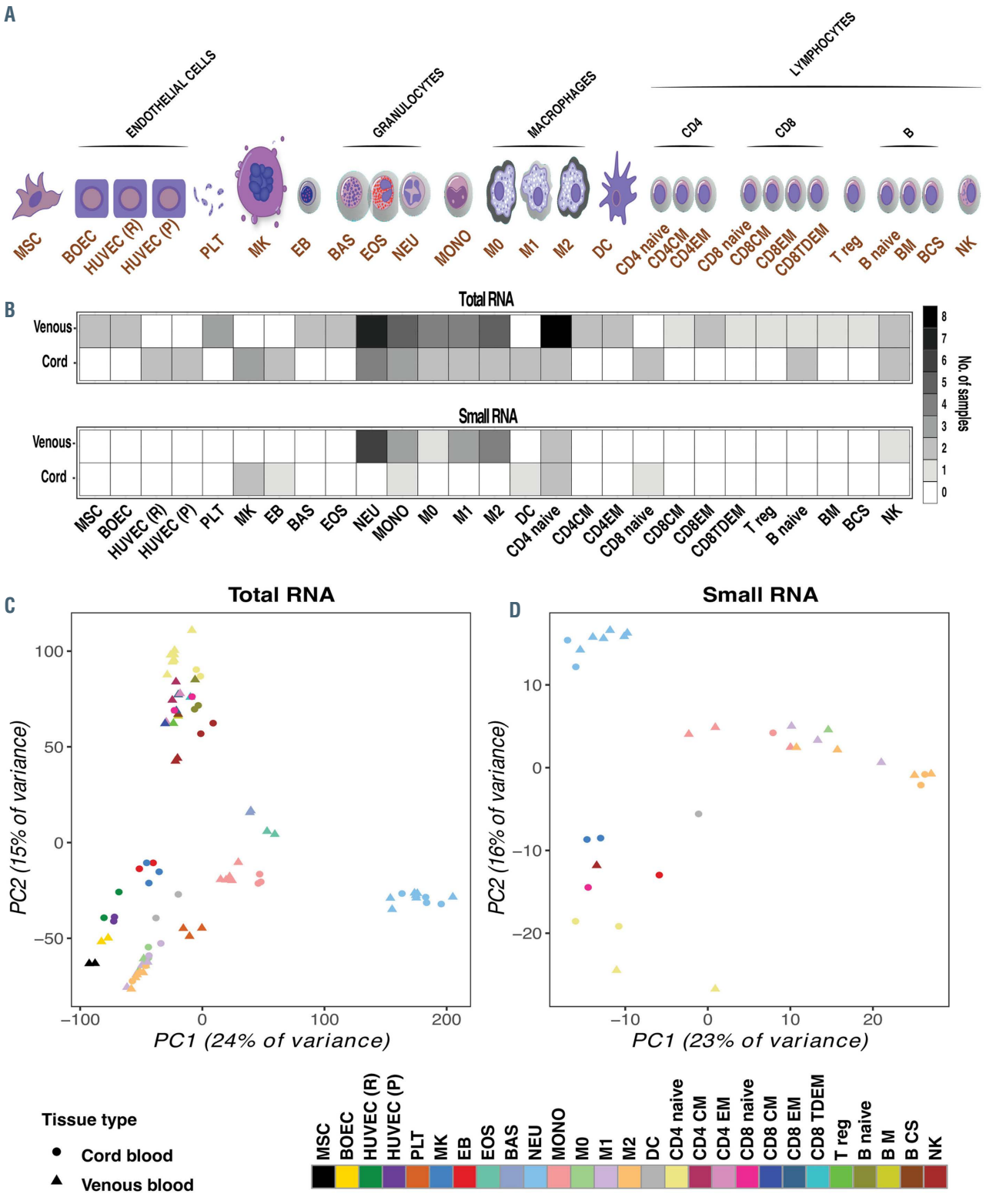

Figure 1. Dataset description and principal component analysis of total and small RNA expression. (A) Graphical representation of the cell types included in the dataset. (B) Heatmaps of the number of samples for each cell type used for total RNA (above) and small RNA (below) sequencing. (C) Scatterplot of the first (PC1) *versus* the second (PC2) principal component of the expression of genes with a log expression estimate greater than zero in at least one sample. (D) Scatterplot of PC1 *versus* PC2 of the expression of the various small RNA with a unique read count >10 in at least one sample. MSC: mesenchymal stem cells; BOEC: blood outgrowth endothelial cell progenitors; HUVEC (R): resting human umbilical vein endothelial cells; HUVEC (P): proliferating human umbilical vein endothelial cells; PLT: platelets; MK: megakaryocytes; EB: erythroblast; BAS: basophils; EOS: eosinophils; NEU: neutrophils; MONO: monocytes; M0: macrophages; M1: lipopolysaccharide-activated macrophages; M2: alternatively activated macrophages; DC: dendritic cells; CD4 naïve: naïve CD4 lymphocytes; CD4CM: central memory CD4 lymphocytes; CD4EM: effector memory CD4 lymphocytes; CD8 naïve: naive CD8 lymphocytes; CD8CM: central memory CD8 lymphocytes; CD8EM: effector memory CD8 lymphocytes; CD8TDEM: terminally differentiated effector memory CD8 lymphocytes; T reg: regulatory CD4 lymphocytes; B naïve: naive B lymphocytes; BM: memory B lymphocytes; BCS: class switch B lymphocytes; NK: natural killer lymphocytes.

reflecting the primary functions corresponding to all cell type categories (*Online Supplementary Table S2*) except BAS, M0 and MONO, at a family-wise error rate <5% (*Online Supplementary Table S7*). Figure 3C illustrates the results of the enrichment analysis for the MK/PLT and DC categories.

Differential expression of microRNA

We applied the same differential expression modeling described above to the small RNA data for which biological replicates were available (MK, NEU, MONO, M1, M2 and CD4TC samples). Of 2,588 miRBase-annotated[26] miRNA, 603 had a posterior probability of differential expression >0.8, of which 573 were classified as cell type-specific. The mean expression of miRNA was strongly associated with having at least one validated target among the 29,920 validated miRNA-messenger RNA (mRNA) interactions in the mirecords, mirtarbase and tarbase databases[27] ($P<2 \times 10^{-16}$, effect size = 0.16, logistic regression). For example, 46 of the 50 miRNA (92%) having the greatest mean expression had at least one validated target, while only 458 (18.2%) of the remaining 2,508 miRNA had a validated target. The miRNA with the greatest expression in their assigned cell type (*Online Supplementary Table S8*) have been previously linked to relevant cellular functions in that cell type. For example, hsa-miR-21-5p (the most highly expressed M1-specific miRNA) is involved in resolution of wound inflammation[28] and macrophage polarization;[9] hsa-let-7g-5p, hsa-miR-26a-5p, hsa-miR-150-5p and hsa-miR-146b-5p (the most highly expressed CD4TC-specific miRNA) are important modulators of $CD4^+$ T cells;[30,31] and hsa-miR-126-3p (the most highly expressed MK-specific miRNA) plays a role in MK/PLT biogenesis.[32,33] However, using the existing databases of miRNA-mRNA interactions, we did not find any correlation between the expression of miRNA and the expression of their targets, which is consistent with miRNA being only one of a diverse set of molecular players in transcriptional regulation of hematopoietic cells and is in agreement with the results of other studies showing that miRNA induce translational repression without mRNA destabilization.[34]

De novo transcriptome assembly identifies novel long non-coding RNA

Several studies have shown that almost two-thirds of the genome is pervasively transcribed,[35] mostly because of the transcription of various types of unannotated non-coding RNA (ncRNA).[36] Among the ncRNA, lncRNA comprise a heterogeneous class of single or multi-exon RNA genes, with crucial roles in controlling gene expression during developmental and differentiation processes.[37] The proportion of RNA species encoded in a genome which are of the lncRNA type increases with developmental complexity, hinting at the importance of RNA-based control mechanisms in the evolution of multicellular organisms.[38] To identify novel transcripts, we assembled sample-specific transcriptomes from read alignments to the reference genome using guided transcriptome assembly,[39] which we then merged into a consensus transcriptome. To avoid the assembly of artefactual sequences originating from pseudogenes, we used a conservative approach that filtered out intronless transcripts and transcripts intersecting any of the transcripts present in Ensembl 75, GENCODE 19 or RefSeq[40] (*Online Supplementary File 3*). This unified filtered transcriptome contained 645 multi-exonic transcripts originating from 400 novel genes. Using the expression values of the subset of 368 novel genes having a log expression >0 in at least one sample, we were able to cluster the samples by cell type (Figure 4A), suggesting that these novel genes might play a role either in the determination of cellular identity or in performing cell type-specific functions.

The vast majority (348 out of 400) of the novel multi-exonic genes had a coding potential below the standard CPAT20 threshold (0.364) used to discriminate potentially coding genes from non-coding genes. However, the 52 potentially coding genes had other characteristics suggesting that they were also non-coding. Firstly, the proportion of their nucleotides overlapping transposon-associated

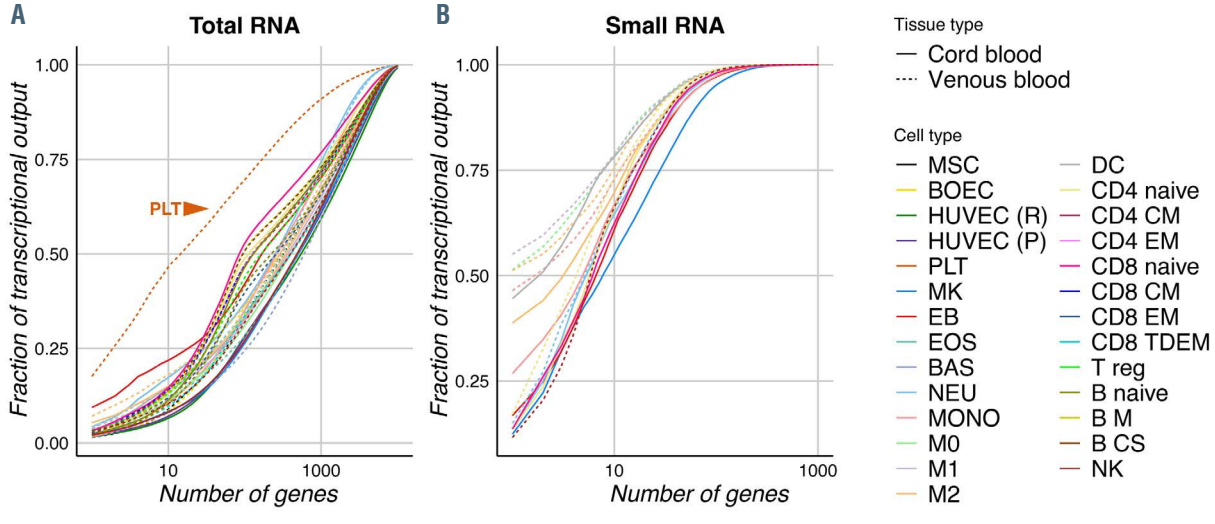

Figure 2. Complexity of genes and miRNA transcriptomes. (A) Cumulative distribution of the fraction of total transcription contributed by non-mitochondrial protein-coding genes when sorted from most to least expressed in each cell type. The x axis is on the $\log_{10}$ scale. (B) Cumulative distribution of the fraction of small RNA transcription contributed by mature miRNA when sorted from most to least expressed in each cell type. The x axis is on the $\log_{10}$ scale. Abbreviations as in Figure 1.

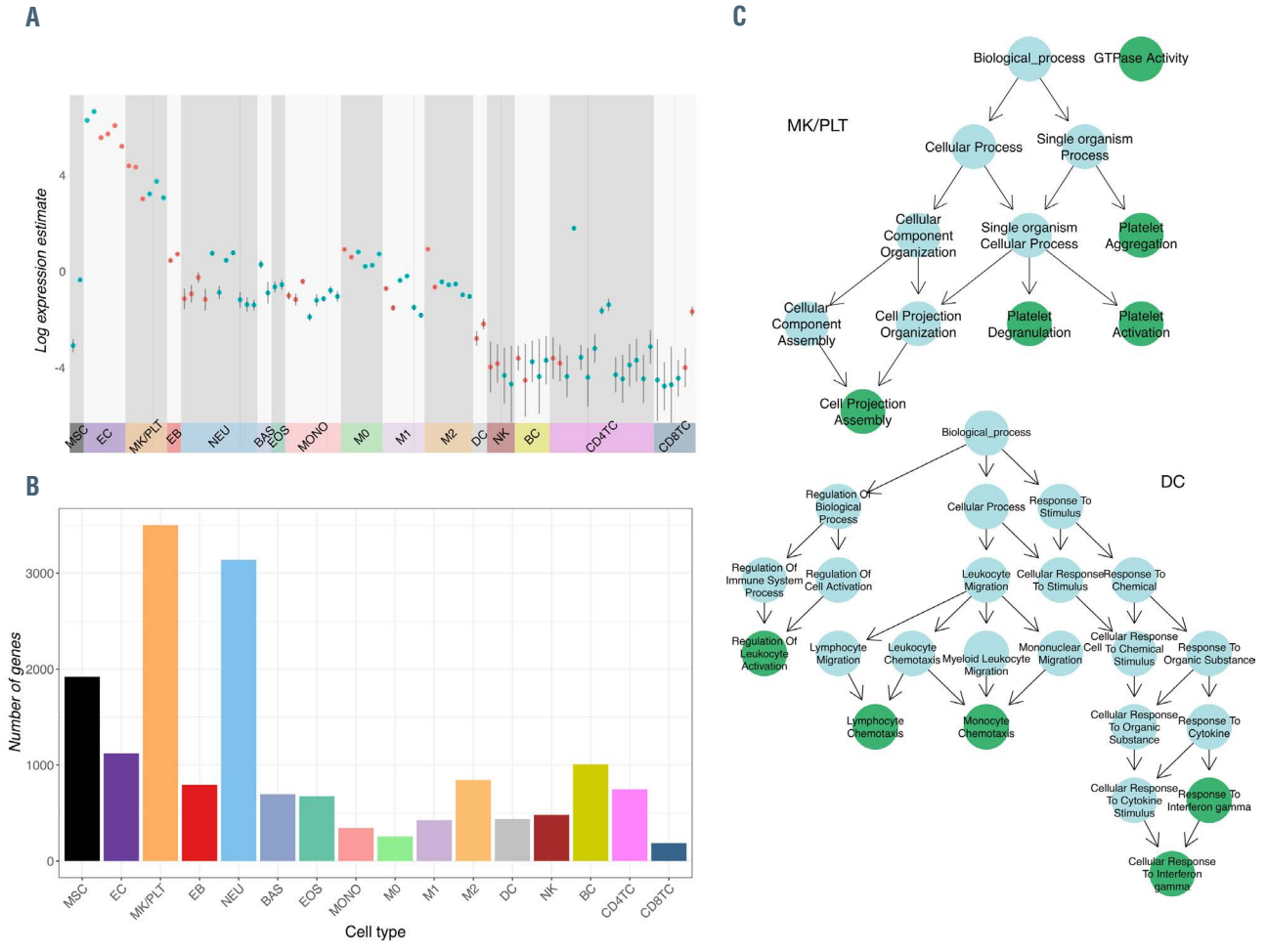

Figure 3. Cell type-specific transcriptional signatures. (A) *VWF* expression estimates and posterior variances. (B) The number of differentially expressed genes classified into each cell type grouping types. (C) Graphical representations of the Gene Ontology term enrichments for the MK/PLT and the DC groups. Note that, as PLT are the immediate anucleated descendants of MK, a gene was assigned to the composite MK/PLT group if it was maximally expressed in either cell type. The nodes represent terms, which are colored green if they are enriched and light blue if they are ontological ancestors of enriched terms, and the edges represent ontological relations. Abbreviations as in Figure 1.

regions and other repetitive or low complexity regions was higher than that of known coding genes and similar to that of novel non-coding genes (*Online Supplementary Figure S3A*). Secondly, their exons had low conservation among vertebrates, with scores resembling those of annotated lncRNA ($P>0.05$, Wilcoxon rank sum test) and novel non-coding genes ($P>0.05$, Wilcoxon rank sum test), and lower than those of protein coding genes ($P\leq0.0001$, Wilcoxon rank sum test) (Figure 4B). Thirdly, their median expression was similar to that of annotated lncRNA and novel genes classified as non-coding by CPAT (median log expression levels: annotated lncRNA, 0.02; novel potentially coding, 0.03; novel non-coding, 0.02; protein-coding genes, 1.2) (Figure 4C). We therefore concluded that all the novel genes, including those with a CPAT score >0.364, were likely to be lncRNA.

Additionally, the novel genes differed from known lncRNA and protein-coding genes in that they had a higher tissue specificity (median Tau: annotated lncRNA, 0.78; novel potentially coding, 0.95; novel non-coding, 0.94; protein-coding genes, 0.49) (Figure 4D). Low expression levels combined with high tissue specificity may explain why these transcripts have not been identified previously. The genomic coordinates of these novel transcripts are provided in *Online Supplementary File 3*.

Circular RNA in mature hematopoietic cells

CircRNA are single stranded RNA molecules of which the ends are covalently joined by a backsplice mechanism. Some circRNA have been shown to regulate transcription[41] or act as miRNA sponges,[42,43] but the majority of circRNA have no known function. Peripheral blood contains thousands of circRNA.[44] We identified backsplice junctions in the total RNA-sequencing dataset using five methods[43-46] and excluded backsplices detected by fewer than three of these methods in order to mitigate methodological biases. In addition, we excluded backsplices overlapping known segmental duplications,[47] multiple genes or Ensembl 75-annotated readthrough transcripts. We thus obtained a list of 91,866 consensus backsplices (*Online Supplementary Table S9*). We further removed junctions observed only in one sample, as they are likely to be spurious, notwithstanding that this may tend to filter junctions specific to cell types with a small number of replicates. In total, 55,187 backsplices were retained for downstream analyses. The majority (81.64%) of these backsplices were exonic and utilized annotated canonical splice sites (Figure 5A), which is consistent with previous reports.[43,48] Almost half (44%) of the backsplices matched structures in circBase[49] exactly and a further 30% shared one of their two splice sites with structures in circBase.

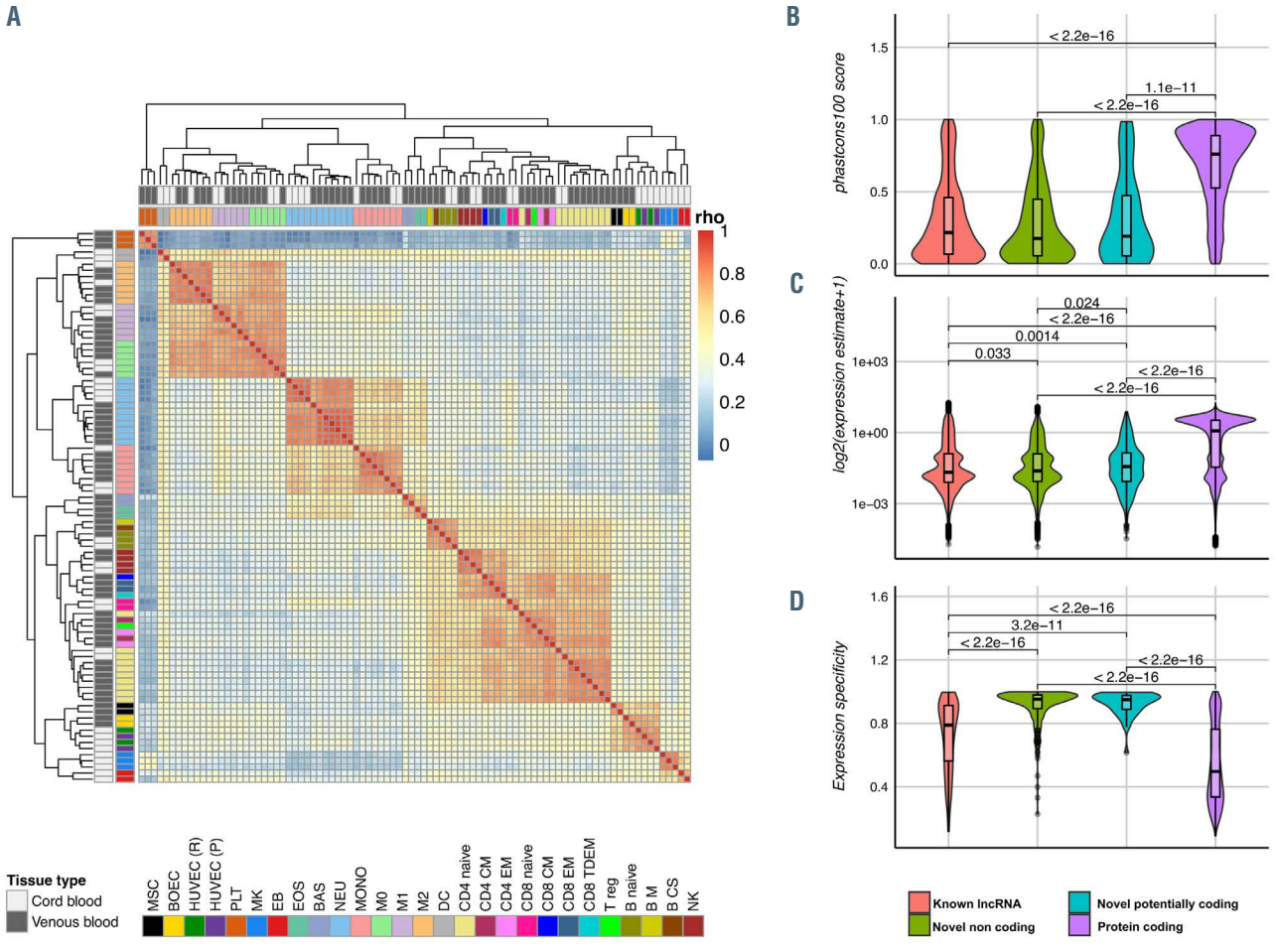

Figure 4. Properties of the identified novel genes. (A) Heatmap of the Spearman rank correlation (rho) matrix computed from the expression estimates of the 368 novel genes expressed (i.e., with a log expression estimate >0) in at least one sample. The dendrogram was computed using complete-linkage clustering with distance specified as one minus the correlation coefficient. (B) Violin plots and overlaid box plots of sequence conservation (UCSC phastCons 100) values for known long non-coding (lnc)RNA, novel non-coding genes, novel potentially coding genes and coding genes annotated in Ensembl 75. The phastCons scores were obtained from multiple alignment of the human (hg19) sequences to the sequences of 99 other vertebrate species. (C) Violin plots and overlaid box plots of expression estimates (expressed as $\log_2(\mu+1)$, where μ is the real scale expression estimate) of known lncRNA, novel non-coding genes, novel potentially coding genes and coding genes annotated in Ensembl 75. (D) Violin plots and overlaid box plots of the expression specificity of known lncRNA, novel non-coding genes, novel potentially coding genes and coding genes annotated in Ensembl 75. (B-D) Pairwise comparisons for which the Wilcoxon signed-rank test yielded $P<0.05$ following Bonferroni adjustment are highlighted. Abbreviations as in Figure 1

In comparison to other RNA species, circRNA have a low rate of formation, but can accumulate inside the cell because they are resistant to exonuclease activity.[50] To investigate the expression patterns of circRNA in hematopoietic cells, we performed hierarchical clustering using Spearman correlations of normalized PTESFinder read counts. This grouped samples by cell type and lineage, showing tissue-specific patterns of circRNA abundance (Figure 5B).

Next, we assessed the variation in the contribution of circRNA abundance to the transcriptional output of each gene. 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 across all genes. We summarized the AP of each cell type as the mean AP over genes and replicates. This cell type-specific summary of AP ranged from 1.02% in resting HUVEC to 12.45% in PLT, which is the only anucleated cell type in our dataset (*Online Supplementary Figure S4A, Online Supplementary Table S10*). Elevated AP in PLT is consistent with the absence of steady-state transcription in PLT and the lower rate of decay of circRNA relative to linear molecules.[51]

We performed differential expression analysis of circRNA between all pairs of functional categories of cell types (*Online Supplementary Table S2*). We identified 5,993 statistically significant differences in circRNA expression, comprising 929 distinct backsplices (<2%) that were differentially expressed in at least one pairwise comparison. These circRNA originated from 698 genes, of which 678 were protein-coding and 20 were non-coding. The maximum number of differentially expressed circRNA in any pairwise comparison was 372 and the median number was 15 (*Online Supplementary File 4*). The expression patterns of differentially expressed circRNA clustered samples by functional category (Figure 5C). To investigate whether the clustering could, in part, be attributed to shared mechanisms of transcription between circRNA and their linear counterparts, we inferred pairwise differential expression of the genes corresponding to the differentially expressed circRNA. There was strong correspondence between the signs of the log fold changes between the

two species of RNA ($P<2 \times 10^{-16}$, odds ratio: 2.31, 95% confidence interval: 2.08–2.57; Fisher exact test) (Figure 5D). Of the 2,122 gene-level comparisons with a posterior probability of differential expression >0.8, over 70% had a log fold change sign in the genes matching that identified in the corresponding circRNA. Although circRNA are typically generated co-transcriptionally, the remainder may reflect cell type-specific competition in their biogenesis with canonical splicing of linear RNA.[52] Several mechanisms of action have been discovered for ncRNA, but only a handful of circRNA have been experimentally verified as functional.[41,43] Furthermore, their functions are distinct from those of their host genes, preventing functional inferences from the analysis of the GO terms of host genes.

Data visualization and download

We have developed a website (*https://blueprint.haem.cam.ac.uk/bloodatlas/*) for generating graphical representations of the data and downloading expression values. Its functionality is showcased in *Online Supplementary Figure S5*.

Discussion

We explored the coding and non-coding transcriptional landscapes of 90 samples comprising 27 different mature hematopoietic cell types (Figure 1A and B). Our aim was to determine how these cell types achieve their unique

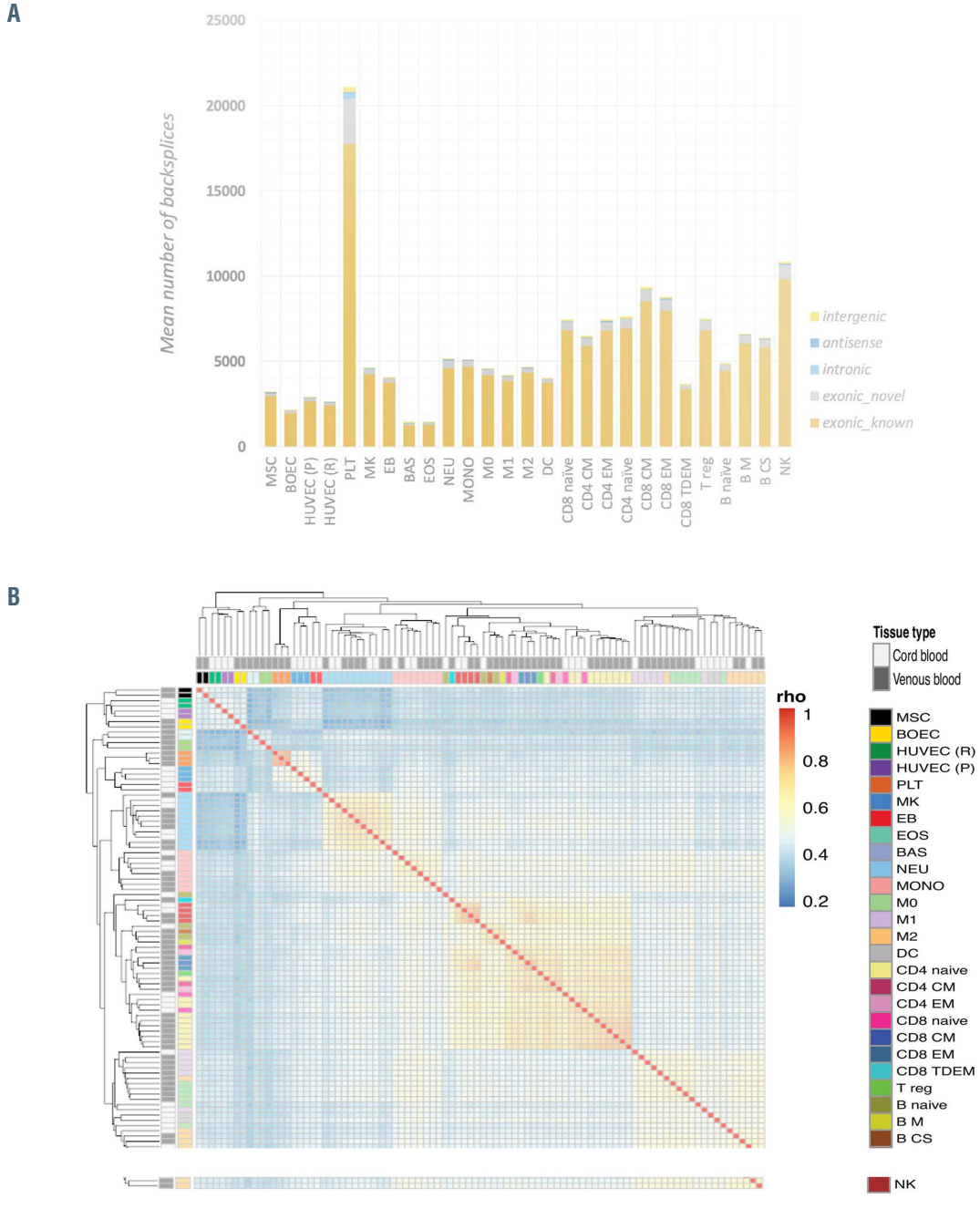

Figure 5. Continued on the following page.

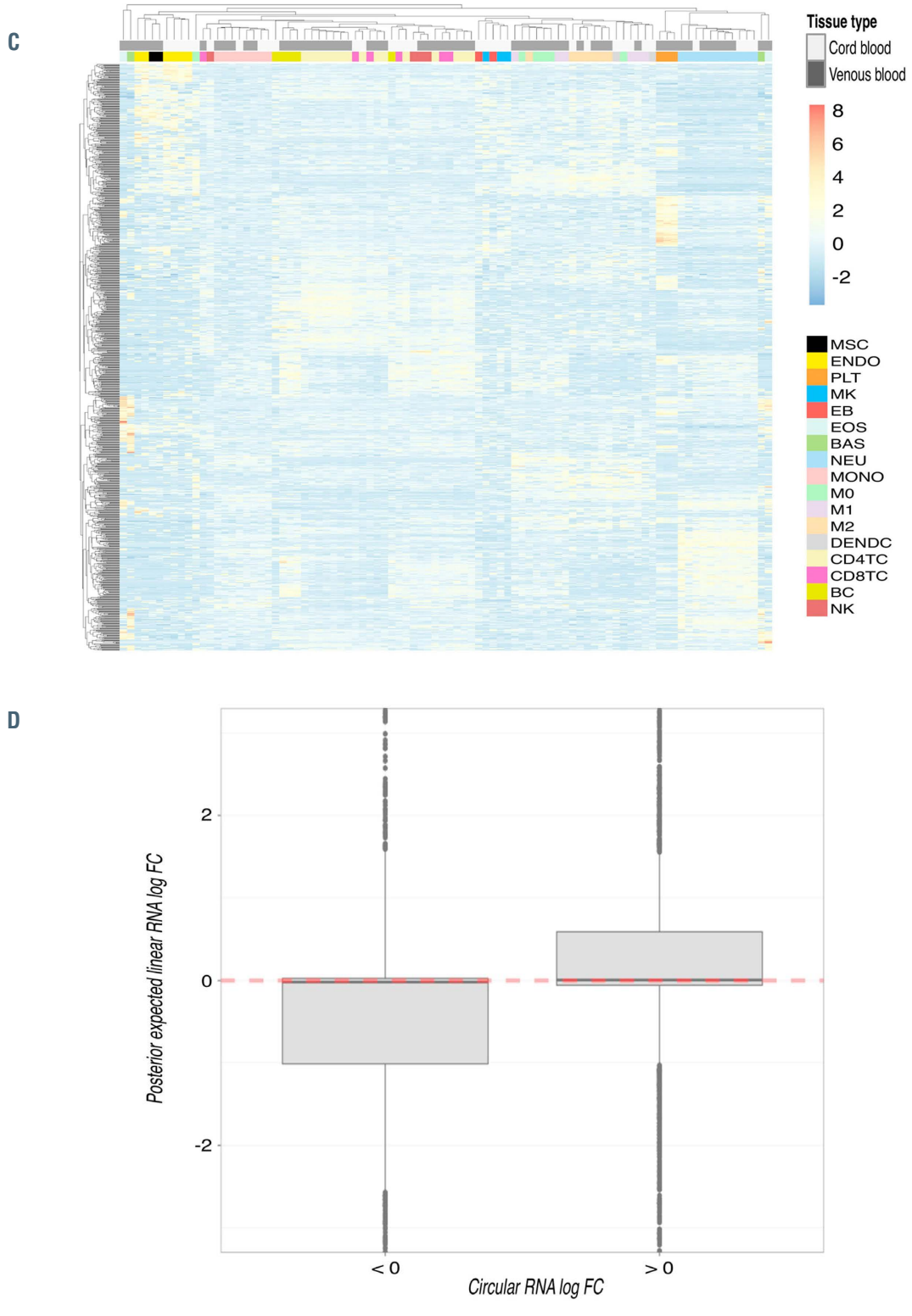

Figure 5. Circular RNA expression in blood cells. (A) Bar plot of the mean number of backsplices identified in each cell type. Each bar is color-coded to indicate the number of backsplices originating in different types of genomic regions: exonic_known: the backsplice corresponds to known splice site; exonic_novel: the backsplice utilizes only one known splice site; intronic: the backsplice is internal to an annotated intron; intergenic: the backsplice junctions do not overlap any annotated exons or introns; antisense: the backsplice is antisense to annotated exons or introns. (B) Heatmap of the Spearman rank correlation (rho) between the backsplice junction counts in each sample. Lowly expressed circular (circ)RNA (having <20 reads in all samples) were excluded. (C) Heatmap of z-scores of the expression estimates for each of the differentially expressed backsplice junctions across cell types. (D) Box plot of the posterior expected log fold-change of the genes corresponding to the significantly differentially expressed circRNA, stratified by the sign of the circRNA log fold change. The posterior expected log fold changes were computed as the log fold changes conditional on differential expression multiplied by the corresponding posterior probabilities of differential expression. The center mark and lower and upper hinges of the boxplots indicate, respectively, the median, 25th and 75th percentiles. Outliers beyond 1.5 times the interquartile range from each hinge are shown. The y-axis covers the range (-3,3). Abbreviations as in Figure 1. FC: fold change.

functional roles in the hematopoietic system. We estimated the transcriptome complexity of each cell type, as it had previously been reported that whole blood is one of the least transcriptionally complex tissues.[24] We found that, out of a mean of ~10,000 expressed protein-coding genes, the number accounting for 75% of each transcriptome ranged from 168 in PLT to 2,422 in resting HUVEC. These genes displayed an enrichment for GO terms relating to basic cellular functions, rather than for terms relating to the different functional phenotypes or identities, the only exception being PLT (Figure 2). These findings indicate that the genes allowing each cell type to perform its functions have a wide range of expression values and they form a unique, although partially overlapping, transcriptional signature. They also suggest that basic cellular functions are maintained even in those cell types with an extremely short half-life, such as neutrophils (*Online Supplementary Table S5*).

To identify the unique gene expression signatures of each cell type, we classified genes according to their cell type specificity after grouping the most similar cell types into functional categories because otherwise they would mutually erase their signals (Figure 3, *Online Supplementary Table S2*). Perhaps not surprisingly given the uniqueness of their function in the coagulation process, the largest signature belongs to the MK/PLT category with 3,502 genes. The smallest signature (186 genes) belongs to the CD8TC group largely due to the overlap with the CD4TC group (*Online Supplementary Table S7*). As expected, the identified signatures showed GO-term enrichment corresponding to the core functions of each cell type, with the exception of BAS, M0 and MONO. This is likely due to the considerable overlaps between the gene expression programs in many of these cell types, which causes the genes to which the primary functions of these cells are ascribed to, not to be selected. Overall, we found that almost 60% of known genes are differentially expressed in the hematopoietic system. Half of these have a high mean expression (log expression >0), whilst only a minority (3.5%) of the non-differentially expressed genes have high mean expression.

The annotation-agnostic nature of RNA-sequencing led us to identify novel genes using guided transcriptome assembly. This approach allowed us to identify 645 multi-exonic novel transcripts from 400 novel genes. The properties of these novel genes, such as their overlap with transposons and repeat elements, low conservation, low expression levels, and high cell type specificity (Figure 4), are in agreement with observations in known lncRNA, as previously shown by Schwarzer and colleagues.[53] The high cell type specificity, in particular, most likely explains why these transcripts have not been identified previously using more coarsely fractionated samples. Moreover, the nature of the library preparation (ribo-depletion, independent of poly-A tail) allowed us to expand the catalog of circRNA transcribed in blood and show that these ncRNA display high levels of cell type specificity (Figure 5). Our findings support the notion that some lncRNA and circRNA may have roles in determining cell fate and functions in hematopoiesis,[53,54] in line with findings in other tissues and organs.[55] However, further work is needed to understand the underlying mechanisms. Finally, to allow a wider access to these data, we created a web-based application (*https://blueprint.haem.cam.ac.uk/bloodatlas/*). Here, expression values at gene and transcript levels, as well as, expression values for microRNA, novel genes and circRNA can be visualized. Moreover, publication-ready graphical representations, together with expression values can also be downloaded.

Disclosures

PF is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. None of the other authors has any conflicts of interest to declare.

Contributions

LG, OGI, NANJ, DS, RP, MK FP and ET performed analyses; MB, FB, SF, NF, JJL, SR, EMR and KD collected samples and generated data; FJM, AF, JMM, LC and PF provided data infrastructure; XE, PF, JHAM, MLY, HGS, WHO, PF and MF provided funding and infrastructure; FJM, AF, PF, FP, ET and MF supervised analyses; LG, OGI, PF, ET and MF conceptualised the analyses.

Acknowledgments

The authors would like to acknowledge the participation of National Institute of Health Research (NIHR) Cambridge BioResource volunteers and thank the NIHR Cambridge BioResource staff for their support.

Funding

The work was funded by a grant from the European Commission 7th Framework Program (FP7/2007–2013, grant 282510, BLUEPRINT) to XE, PF, JHAM, MY, HGS and WHO. WHO is an NIHR senior investigator and receives funding from Bristol-Myers Squibb, the British Heart Foundation, the Medical Research Council and the NIHR. OGI, FJM, AF, JMM, LC and PF are funded by the Wellcome Trust (WT108749/Z/15/Z) with additional funding for specific project components such as GENCODE from the National Human Genome Research Institute of the National Institutes of Health (2U41HG007234). The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. KD is a HSST trainee supported by NHS Health Education England. NF is funded by the NIHR Cambridge Biomedical Research Centre. FP is supported by the Fundação Carlos Chagas Filho de Amparo à Pesquisado Estado do Rio de Janeiro (FAPERJ; E-26/203.229/2016). NANJ is a recipient of a scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001). The work by DS was supported in part by an Isaac Newton fellowship to MF. MF is supported by the British Heart Foundation (FS/18/53/33863).

References

1. Bagger FO, Sasivarevic D, Sohi SH, et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. Nucleic Acids Res. 2016;44(D1):D917-24.
2. Novershtern N, Subramanian A, Lawton LN, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011;144(2):296-309.
3. Laurenti E, Doulatov S, Zandi S, et al. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat Immunol. 2013;14(7):756-763.
4. Caron H, van Schaik B, van der Mee M, et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science. 2001;291(5507):1289-1292.

5. Kapranov P, Cawley SE, Drenkow J, et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002; 296(5569):916-919.
6. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133-141.
7. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290-295.
8. Adams D, Altucci L, Antonarakis SE, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012; 30(3):224-226.
9. Stunnenberg HG; International Human Epigenome Consortium, Hirst M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;1 67(5):1145-1149.
10. Chen L, Kostadima M, Martens JHA, et al. Transcriptional diversity during lineage commitment of human blood progenitors. Science. 2014;345(6204):1251033.
11. Chen L, Ge B, Casale FP, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167(5):1398-1414.
12. Turro E, Su SY, Goncalves A, et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 2011;12(2):R13.
13. Turro E, Astle WJ, Tavare S. Flexible analysis of RNA-seq data using mixed effects models. Bioinformatics. 2014;30(2):180-188.
14. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
15. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078-2079.
16. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7(3):562-578.
17. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-1774.
18. Rosenbloom KR, Armstrong J, Barber GP, et al. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 2015;43 (Database issue):D670-681.
19. Lawrence M, Huber W, Pages H, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013; 9(8):e1003118.
20. Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6):e74.
21. Yanai I, Benjamin H, Shmoish M, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21(5):650-659.
22. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-1050.
23. Genotype-Tissue Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-585.
24. Mele M, Ferreira PG, Reverter F, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348(6235):660-665.
25. Mercer TR, Neph S, Dinger ME, et al. The human mitochondrial transcriptome. Cell. 2011;146(4):645-658.
26. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68-73.
27. Ru Y, Kechris KJ, Tabakoff B, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic Acids Res. 2014;42(17):e133.
28. Das A, Ganesh K, Khanna S, Sen CK, Roy S. Engulfment of apoptotic cells by macrophages: a role of microRNA-21 in the resolution of wound inflammation. J Immunol. 2014;192(3):1120-1129.
29. Wang Z, Brandt S, Medeiros A, et al. MicroRNA 21 is a homeostatic regulator of macrophage polarization and prevents prostaglandin E2-mediated M2 generation. PLoS One. 2015;10(2):e0115855.
30. Yu HR, Hsu TY, Huang HC, et al. Comparison of the functional microRNA expression in immune cell subsets of neonates and adults. Front Immunol. 2016; 7:615.
31. Ghisi M, Corradin A, Basso K, et al. Modulation of microRNA expression in human T-cell development: targeting of NOTCH3 by miR-150. Blood. 2011; 117(26):7053-7062.
32. Opalinska JB, Bersenev A, Zhang Z, et al. MicroRNA expression in maturing murine megakaryocytes. Blood. 2010;116(23):e128-138.
33. Ple H, Landry P, Benham A, et al. The repertoire and features of human platelet microRNAs. PLoS One. 2012;7(12):e50746.
34. Bazzini AA, Lee MT, Giraldez AJ. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science. 2012;336(6078):233-237.
35. Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005;308(5725):1149-1154.
36. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799-816.
37. Flynn RA, Chang HY. Long noncoding RNAs in cell-fate programming and reprogramming. Cell Stem Cell. 2014;14(6):752-761.
38. Taft RJ, Pheasant M, Mattick JS. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays. 2007;29(3):288-299.
39. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
40. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733-745.
41. Li Z, Huang C, Bao C, et al. Exon-intron circular RNAs regulate transcription in the nucleus. Nat Struct Mol Biol. 2015;22(3):256-264.
42. Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. Nature. 2013; 495(7441):384-388.
43. Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495(7441):333-338.
44. Westholm JO, Miura P, Olson S, et al. Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell Rep. 2014;9(5):1966-1980.
45. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. 2015; 16:4.
46. Zhang XO, Dong R, Zhang Y, et al. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. 2016;26(9):1277-1287.
47. 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;11(6):1005-1017.
48. Starke S, Jost I, Rossbach O, et al. Exon circularization requires canonical splice signals. Cell Rep. 2015;10(1):103-111.
49. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20(11):1666-1670.
50. Rybak-Wolf A, Stottmeister C, Glazar P, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell. 2015;58(5):870-885.
51. Alhasan AA, Izuogu OG, Al-Balool HH, et al. Circular RNA enrichment in platelets is a signature of transcriptome degradation. Blood. 2016;127(9):e1-e11.
52. Ashwal-Fluss R, Meyer M, Pamudurti NR, et al. circRNA biogenesis competes with pre-mRNA splicing. Mol Cell. 2014;56(1):55-66.
53. Schwarzer A, Emmrich S, Schmidt F, et al. The non-coding RNA landscape of human hematopoiesis and leukemia. Nat Commun. 2017;8(1):218.
54. Alvarez-Dominguez JR, Lodish HF. Emerging mechanisms of long noncoding RNA function during normal and malignant hematopoiesis. Blood. 2017;130(18):1965-1975.
55. Lorenzi L, Chiu H-S, Cobos FA, et al. The RNA Atlas, a single nucleotide resolution map of the human transcriptome. bioRxiv. 2019 Oct 17. [Epub ahead of print].