Cell type specific novel IncRNAs and circRNAs in the BLUEPRINT haematopoietic transcriptomes atlas

by Luigi Grassi, Osagie G. Izuogu, Natasha A.N. Jorge, Denis Seyres, Mariona Bustamante, Frances Burden, Samantha Farrow, Neda Farahi, Fergal J. Martin, Adam Frankish, Jonathan M. Mudge, Myrto Kostadima, Romina Petersen, John J. Lambourne, Sophia Rowlston, Enca Martin-Rendon, Laura Clarke, Kate Downes, Xavier Estivill, Paul Flicek, Joost H.A. Martens, Marie-Laure Yaspo, Hendrik G. Stunnenberg, Willem H. Ouwehand, Fabio Passetti, Ernest Turro, and Mattia Frontini

Haematologica 2020 [Epub ahead of print]

Citation: Luigi Grassi, Osagie G. Izuogu, Natasha A.N. Jorge, Denis Seyres, Mariona Bustamante, Frances Burden, Samantha Farrow, Neda Farahi, Fergal J. Martin, Adam Frankish, Jonathan M. Mudge, Myrto Kostadima, Romina Petersen, John J. Lambourne, Sophia Rowlston, Enca Martin-Rendon, Laura Clarke, Kate Downes, Xavier Estivill, Paul Flicek, Joost H.A. Martens, Marie-Laure Yaspo, Hendrik G. Stunnenberg, Willem H. Ouwehand, Fabio Passetti, Ernest Turro and Mattia Frontini. Cell type specific novel IncRNAs and circRNAs in the BLUEPRINT haematopoietic transcriptomes atlas. Haematologica. 2020; 105:xxx

Publisher's Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors' final approval; the final version of the manuscript will then appear in print on a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
Cell type specific novel IncRNAs and circRNAs in the BLUEPRINT haematopoietic 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 Rowston$^{1,2}$, 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 CB2 0PT, UK. 2, National Health Service Blood and Transplant, Cambridge Biomedical Campus, Cambridge CB2 0PT, UK. 3, National Institute for Health Research BioResource, Rare Diseases, Cambridge University Hospitals, Cambridge, United Kingdom. 4, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. 5, Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz, Rio de Janeiro 21040-360, 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, United Kingdom. 10, R&D Division, National Health Service (NHS)-Blood and Transplant, Oxford Centre, Oxford, United Kingdom. 11, Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, United Kingdom. 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 CB10 1HH, UK. 16, British Heart Foundation Centre of Excellence, Cambridge Biomedical Campus, Long Road, Cambridge CB2 0QQ, UK. 17, Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz, Curitiba 81350-010, Brazil. 18, Medical Research Council Biostatistics Unit, Cambridge Institute of Public Health, Cambridge Biomedical Campus, Cambridge, UK. 19, Institute of Biomedical & Clinical Science, College of Medicine and Health, University of Exeter Medical School, RILD Building (Level 4), Barrack Road, Exeter, EX2 5DW. ∑ present address: Biopharmaceutical Development, BioPharmaceuticals R&D, AstraZeneca, Granta Park, Cambridge, CB21 6GH, UK. * Joint first authors. § Corresponding authors: ET et341@cam.ac.uk, MF mf471@cam.ac.uk or m.frontini@exeter.ac.uk

Abstract

Transcriptional profiling of hematopoietic cell subpopulations has helped characterize the developmental stages of the hematopoietic system and the molecular bases of malignant and non-malignant blood diseases for the past three decades. Previously, only the genes targeted by expression microarrays could be profiled genome wide. High-throughput RNA sequencing (RNA-seq), however,
encompasses a broader repertoire of RNA molecules, without restriction to previously annotated genes. We analysed the BLUEPRINT consortium RNA-seq data for mature hematopoietic cell types. The data comprised 90 total RNA-seq samples, each composed of one of 27 cell types, and 32 small RNA-seq 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 RNAs 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 hematological development and diseases. Finally, we made the data accessible via a web-based interface: https://blueprint.haem.cam.ac.uk/bloodatlas/.

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 hematological 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 have been performed on increasingly purified primary hematopoietic cell populations in the post genome era\textsuperscript{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\textsuperscript{4}, despite the early discovery that transcription is widespread throughout the human genome\textsuperscript{5}. The introduction of high-throughput nucleic acid sequencing technologies\textsuperscript{6} has improved the assembly of the human genome and the annotation of transcriptomes therein, and it has enabled a more comprehensive analysis of gene expression using transcriptomic assembly approaches\textsuperscript{7}. The BLUEPRINT consortium\textsuperscript{8} was established to characterize the epigenetic state and transcriptional profile of different types of hematopoietic cell. Reference datasets for DNA methylation, histone modifications and gene expression were generated using state-of-the-art technologies from highly purified cell populations, in accordance with quality standards set by the International Human Epigenome Consortium\textsuperscript{9}. RNA-seq data from over 270 samples encompassing 55 cell types have been made publicly available (http://dcc.blueprint-epigenome.eu), a subset of which has been described previously\textsuperscript{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\textsuperscript{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\textsuperscript{7} using total RNA-seq 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 RNAs (circRNAs), 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 normalised expression data (https://blueprint.haem.cam.ac.uk/bloodatlas/).

Methods

Ethical approval
Samples were obtained from NHS 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
Please refer to the Supplementary Material for the protocols used for cell isolations, RNA extraction and library construction.

Bioinformatic analysis
An overview of the bioinformatic pipeline is shown in Fig. S1. To analyse 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\textsuperscript{14} (1.0.1; parameters “-a --best --strata -S -m 100 -X 500 --chunkmbs 256 --nofw --fr”). Small RNA-seq 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) (30423142) human mature miRNAs with RapMap (v 0.4.0) (27307617) using the parameters “quasimap -c -s -z 0.9”. We used MMSEQ\textsuperscript{12} and MMDIFF\textsuperscript{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 modelling 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)\textsuperscript{15}. We performed guided transcriptome assembly for each sample using StringTie\textsuperscript{7} (v 1.3.4; parameters “-p 8 --rf -G Ensembl_75.gtf -v -I 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\textsuperscript{16}. We identified intergenic transcripts and filtered out the ones overlapping known transcripts annotated in Gencode (v19)\textsuperscript{17} and UCSC (v hg19)\textsuperscript{18}
using the GenomicRanges package\textsuperscript{19}. We assessed the protein coding potential of the novel intergenic multi-exonic transcripts using CPAT (v 1.2.4)\textsuperscript{20}. We chose CPAT due to its superior accuracy relative to competing methods\textsuperscript{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 \( \text{Tau} \)\textsuperscript{21} to compare the cell type specificities of novel genes, known lncRNAs and known protein coding genes. We used the BioConductor R package “phastCons100way.UCSC.hg19”\textsuperscript{22} to obtain sequence conservation scores of novel genes, known lncRNAs and known protein coding genes. A detailed description of the computational methods used to identify circRNAs, compare their sequences to known sequences and quantify expression levels is given in the 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/EGAC000010001351). The dataset IDs are listed in 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 (Fig. 1A, Fig. 1B and Table S1) from 72 whole blood and cord blood donations, either by magnetic bead separation or by flow activated cell sorting (see M&M). 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), LPS activated macrophages (M1), alternatively activated macrophages (M2), dendritic cells (DC), naive CD4 lymphocytes, central memory CD4 lymphocytes (CD4CM), effector memory CD4 lymphocytes (CD4EM), regulatory CD4 lymphocytes (TREG), naive CD8 lymphocytes, central memory CD8 lymphocytes (CD8 CM), effector memory CD8 lymphocytes (CD8 EM), terminally differentiated effector memory CD8 lymphocytes (CD8 TDEM), naive B lymphocytes, memory B lymphocytes (BM), 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-seq is presented in Figs. 1A and 1B and Table S2. We generated a mean of 91M 75bp 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 150bp paired-end reads (Table S1). We also generated a mean of 4.5M 50bp single end reads for small RNA samples (Table S3). Principal
component analysis (PCA) of the log expression estimates for both protein-coding genes and small RNAs 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 (Fig. 1C, 1D, S2A and S2B). This correspondence was also apparent by hierarchical clustering of samples using Spearman's rank correlation (Fig. S2C and S2D).

The GTEx project\(^{23}\) showed that whole blood has a very low gene expression complexity compared to 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 analysed transcriptome complexity in different blood cell types. 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 human umbilical vein endothelial cells (HUVEC (R); Fig. 2A, Table S4, Supplementary File 1). With the exception of PLT, the sets of genes yielding 75% of total expression in each cell type showed gene ontology (GO) terms enrichment 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 (Table S5). The corresponding analysis of the small RNA data showed a very low complexity: between 1 and 7 miRNAs accounted for 50% of total expression and fewer than 10 miRNAs accounted for 75% of the expression in each of the 11 cell types (Fig. 2B, 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 (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 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 haematopoiesis is a few hundred. The differentially expressed genes were then classified by the cell type with the greatest expression. 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 (M&M, Table S6). For example, VWF was assigned the endothelial cells label (EC) because, firstly, its expression varies across cell types (posterior probability of differential expression ~ = 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 (Fig. 3A). The number of genes assigned to each category ranged from 186 in CD8 T lymphocytes (CD8TC) to 3,502 in MK/PLT (Fig. 3B). Using these groups of genes, we found enrichment for GO terms reflecting the primary functions corresponding to all cell type categories (Table S2) except BAS, M0 and MONO, at a family-wise error rate < 5% (Table S7). Fig. 3C illustrates the results of the enrichment analysis for the MK/PLT and DC categories.

**Differential expression of miRNAs.**

We applied the same differential expression modelling described above to the small RNA data for which biological replicates were available (MK, NEU, MONO, M1, M2 and CD4TC samples). 603 out of 2,588 miRBase-annotated26 miRNAs had a posterior probability of differential expression >0.8, of which 573 were classified as cell type specific. The mean expression of miRNAs was strongly associated with having at least one validated target amongst the 29,920 validated miRNA-mRNA interactions in the mirecords, mirtarbase and tarbase databases 27 (P < 2 x 10^-16, effect size = 0.16, logistic regression). For example, 46 of the 50 miRNAs (92%) having the greatest mean expression had at least one validated target, while only 458/2508 (18.2%) of the remaining 2,508 miRNAs had a validated target. The miRNAs with the greatest expression in their assigned cell type (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 29; hsa-let-7g-5p, hsa-miR-26a-5p, hsa-miR-150-5p and hsa-miR-146b-5p (the most highly expressed CD4TC-specific miRNAs) 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. Using the existing databases of miRNA-mRNA interactions, however, we did not find any correlation between the expression of miRNAs and the expression of their targets, which is consistent with miRNAs being only one of a
diverse set of molecular players in transcriptional regulation of haematopoietic cells and is in agreement with the results of other studies showing that miRNAs induce translational repression without mRNA destabilisation\textsuperscript{34}.

**De novo transcriptome assembly identifies novel lncRNAs.**

Several studies have shown that almost two thirds of the genome is pervasively transcribed\textsuperscript{35}, mostly due to the transcription of various types of unannotated non-coding RNAs (ncRNAs)\textsuperscript{36}. Among the ncRNAs, lncRNAs comprise a heterogeneous class of single or multi-exon RNA genes, with crucial roles in controlling gene expression during developmental and differentiation processes\textsuperscript{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\textsuperscript{38}. To identify novel transcripts, we assembled sample-specific transcriptomes from read alignments to the reference genome using guided transcriptome assembly\textsuperscript{39}, which we then merged into a consensus transcriptome. To avoid the assembly of artefactual sequences originating from pseudo-genes, 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\textsuperscript{40} (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 (Fig. 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 Coding-Potential Assessment Tool (CPAT)\textsuperscript{20} threshold (0.364) used to discriminate potentially coding 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 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 (Fig. S3A). Secondly, their exons had low conservation amongst vertebrates, with scores resembling those of annotated lncRNAs ($P > 0.05$, Wilcoxon rank sum test) and novel non-coding genes ($P > 0.05$, Wilcoxon rank sum test), and lower than those of protein coding genes ($P <= 0.0001$, Wilcoxon rank sum test; Fig. 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; Fig. 4C). We therefore concluded that all the novel genes, including those with a CPAT score > 0.364, were likely to be lncRNAs.
Additionally, the novel genes differed from known IncRNAs and protein coding genes in that they had a higher tissue specificity (median Tau: annotated IncRNA, 0.78; novel potentially coding, 0.95; novel non-coding, 0.94; protein coding genes, 0.49; Fig. 4D). Low expression levels combined with high tissue specificity may explain why these transcripts have not previously been identified. The genomic coordinates of these novel transcripts are provided in Supplementary file 3.

Circular RNA in mature hematopoietic cells.
Circular RNAs (circRNAs) are single stranded RNA molecules for which the ends are covalently joined by a backsplice mechanism. Some circRNAs have been shown to regulate transcription\(^4\) or act as miRNA sponges\(^42, 43\), but the majority of circRNAs have no known function. Peripheral blood contains thousands of circRNAs\(^44\). We identified backsplice junctions in the total RNA-seq 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 (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 (Fig. 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.

In comparison to other RNA species, circRNAs have a low rate of formation, but can accumulate inside the cell because they are resistant to exonuclease activity\(^5\)\(^0\). To investigate the expression patterns of circRNAs in hematopoietic cells, we performed hierarchical clustering using Spearman’s correlations of normalised PTESFinder read counts. This grouped samples by cell type and lineage, showing tissue-specific patterns of circRNA abundance (Fig. 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 summarised 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 HUVEC (R) to 12.45% in PLT, which is the only anucleated cell type in our dataset (Fig. S4A and Table S10). Elevated AP in PLT is consistent with the absence of steady-state transcription in PLT and the lower rate of decay of circRNAs relative to linear molecules\(^51\).
We performed differential expression analysis of circRNAs between all pairs of functional categories of cell types (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 circRNAs originated from 698 genes, of which 678 were protein-coding and 20 were non-coding. The maximum number of differentially expressed circRNAs in any pairwise comparison was 372 and the median number was 15 (Supplementary File 4). The expression patterns of differentially expressed circRNAs clustered samples by functional category (Fig. 5C). To investigate whether the clustering could, in part, be attributed to shared mechanisms of transcription between circRNAs and their linear counterparts, we inferred pairwise differential expression of the genes corresponding to the differentially expressed circRNAs. 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's exact test; Fig. 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 circRNAs are typically generated co-transcriptionally, the remainder may reflect cell-type specific competition in their biogenesis with canonical splicing of linear RNAs\textsuperscript{52}. Several mechanisms of action have been discovered for non-coding RNAs, but only a handful of circRNAs have been experimentally verified as functional.\textsuperscript{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/](https://blueprint.haem.cam.ac.uk/bloodatlas/)) for generating graphical representations of the data and downloading expression values. Its functionality is showcased in Fig. S5.

**Discussion**

We explored the coding and non-coding transcriptional landscapes of 90 samples comprising 27 different mature hematopoietic cell types (Fig. 1A and 1B). Our aim was to determine how these cell types achieve their unique 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\textsuperscript{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 HUVEC(R). 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 (Fig. 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 up-kept even in those cell types with an
extremely short half-life, such as neutrophils (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 (Fig. 3, 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 3502 genes. The smallest signature (186 genes) belongs to the CD8TC group largely due to the overlap with the CD4TC group (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-seq 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 (Fig. 4), are in agreement with observations in known IncRNAs, as previously shown by Schwarzer and colleagues\(^5\). 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 catalogue of circRNAs transcribed in blood and show that these ncRNAs display high levels of cell type specificity (Fig. 5). Our findings support the notion that some IncRNAs and circRNAs may have roles in determining cell fate and functions in hematopoiesis\(^5\),\(^5\), in line with findings in other tissues and organs\(^5\). 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 level, as well as, expression values for microRNAs, novel genes and circRNAs can be visualized. Moreover, publication-ready graphical representations, together with expression values can also be downloaded.

Acknowledgments and funding.
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. 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), accordingly 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 Pesquisa do 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). DS work has been supported in part by an Isaac Newton fellowship to MF. MF is supported by the British Heart Foundation (FS/18/53/33863).

Conflict of interest.
PF is a member of the scientific advisory boards of Fabric Genomics, Inc., and Eagle Genomics, Ltd. All other authors have no CoI to declare.

REFERENCES


Figures legends

Figure 1: Dataset description and principal component analysis of total and small RNA expression.

1A: Graphical representation of the cell types included in the dataset. Here and throughout, the following acronyms are used: MSC, mesenchymal stem cells; BOEC, blood outgrowth endothelial cell progenitors; HUVEC (R) and (P), umbilical vein endothelial cells, resting and proliferating, respectively; PLT, platelets; MK, megakaryocytes; EB, erythroblast; BAS, basophils; EOS, eosinophils; NEU, neutrophils; MONO, monocytes; M0, macrophages; M1, LPS activated macrophages; M2, alternatively activated macrophages; DC, dendritic cells; CD4 naive, naive CD4 lymphocytes; CD4CM, central memory CD4 lymphocytes; CD4EM, effector memory CD4 lymphocytes; CD8 naive, 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 naive, naive B lymphocytes; BM, memory B lymphocytes; BCS, class switch B lymphocytes; NK, natural killer lymphocytes. 1B: Heatmaps of the number of samples for each cell type used for total RNA (above) and small RNA (below) sequencing. 1C: Scatterplot of the first (PC1) vs the second (PC2) principal component of the expression of genes with a log expression estimate greater than zero in at least one sample. 1D: Scatterplot of PC1 vs PC2 of the expression of the small RNAs with a unique read count >10 in at least one sample.

Figure 2: Complexity of genes and miRNA transcriptomes.

2A: 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. 2B: Cumulative distribution of the fraction of small RNA transcription contributed by mature miRNAs when sorted from most to least expressed in each cell type. The x axis is on the log_{10} scale.

Figure 3: Cell type specific transcriptional signatures.

3A: VWF expression estimates and posterior variances. 3B: The number of differentially expressed genes classified into each cell type grouping types. 3C: Graphical representations of the GO 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 coloured green if they are enriched and light blue if they are ontological ancestors of enriched terms, and the edges represent ontological relations.

Figure 4: Properties of the identified novel genes.

4A: Heatmap of Spearman's 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.
4B: Violin plots and overlaid box plots of sequence conservation (UCSC phastCons 100) values for known lncRNAs, 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. 4C: Violin plots and overlaid box plots of expression estimates (expressed as \(\log_2(\mu+1)\), where \(\mu\) is the real scale expression estimate) of known lncRNAs, novel non-coding genes, novel potentially coding genes and coding genes annotated in Ensembl 75. 4D: Violin plots and overlaid box plots of the expression specificity of known lncRNAs, novel non-coding genes, novel potentially coding genes and coding genes annotated in Ensembl 75. 4B–D: Pairwise comparisons for which the Wilcoxon signed-rank test yielded \(P < 0.05\) following Bonferroni adjustment are highlighted.

Figure 5: CircRNA expression in blood cells.
5A: Bar plot of the mean number of backsplices identified in each cell type. Each bar is colour-coded to indicate the number of backsplices originating in different types of genomic region: exonic_known, the backsplice corresponds to known splice site; exonic_novel, the backsplice utilises 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. 5B: Heatmap of the Spearman’s rank correlation (rho) between the backsplice junction counts in each sample. Lowly expressed circRNAs (having <20 reads in all samples) were excluded. 5C: Heatmap of z-scores of the expression estimates for each of the differentially expressed backsplice junctions across cell types. 5D: Box plot of the posterior expected log fold change of the genes corresponding to the significantly differentially expressed circular RNAs, stratified by the sign of the circular RNA 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 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. The y-axis covers the range (-3,3).
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 extensively1-4. 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 described5. 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 described6. 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\textsuperscript{14}. Z-scores for differentially expressed backsplices identified by DESeq2 were computed over samples from the normalised backsplice read counts.
**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 IncRNAs, 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.
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 as Excel file only

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 as Excel file only

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


