Title: Transcriptional diversity during lineage commitment of human blood progenitors

Authors: Lu Chen[1,2,3†], Myrto Kostadima[2,4,3†], Joost H.A. Martens[5†], Giovanni Canu[2,3], Sara P. Garcia[2,3], Ernest Turro[2,3], Kate Downes[2,3], Iain C. Macaulay[7], Ewa Bielczyk-Maczynska[2,3], Sophia Coe[2,3], Samantha Farrow[2, 3], Pawan Poudel[2,3], Frances Burden[2,3], Sjoert B.G. Jansen[2,3], William J. Astle[2,3,6], Antony Attwood[2,3], Tadbir Bariana[8,9], Bernard de Bono[10,11], Alessandra Breschi[12], John C. Chambers[13,14], BRIDGE Consortium[‡], Fizzah A. Choudry[2,3], Laura Clarke[4], Paul Coupland[1], Martijn van der Ent[5], Wendy N. Erber[15], Joop H. Jansen[16], Rémi Favier[17], Matthew E. Fenech[18], Nicola Foad[2,3], Kathleen Freson[19], Chris van Geet[19], Keith Gomez[9], Roderic Guigo[12], Daniel Hampshire[2,3], Anne M. Kelly[2,3,20], Hindrik H.D. Kerstens[5], Jaspal S. Kooner[13,14], Michael Laffan[21], Claire Lentaigne[21], Charlotte Labalette[2,3], Tiphaine Martin[2,3,22±], Stuart Meacham[2,3], Andrew Mumford[23], Sylvia Nürnberg[2,3§], Emilio Palumbo[12], Bert A. van der Reijden[16], David Richardson[4], Stephen J. Sammut[24,25], Greg Slodkowicz[4], Asif U. Tamuri[4], Louella Vasquez[3], Katrin Voss[2,3¶], Stephen Watt[3], Sarah Westbury[26], Paul Flicek[4,1], Remco Loos[4], Nick Goldman[4], Paul Bertone[4,27,28], Randy J. Read[29], Sylvia Richardson[6], Ana Cvejic[2,1], Nicole Soranzo[1,2†], Willem H. Ouwehand[2,3,1†], Hendrik G. Stunnenberg[5†], Mattia Frontini[2,3†*], Augusto Rendon[2,3,6†*]

Affiliations:

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

[2]Department of Haematology, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[3]NHS Blood and Transplant, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom.

[5]Department of Molecular Biology, Radboud University, Nijmegen, the Netherlands.

[6]Medical Research Council Biostatistics Unit, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[7]Sanger Institute-EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom

[8]Department of Haematology, University College London Cancer Institute, London, United Kingdom.

[9]The Katharine Dormandy Haemophilia Centre and Thrombosis Unit, Royal Free NHS Trust, London, United Kingdom.

[10]CHIME Institute, University College London, Archway Campus, London, United Kingdom.

[11]Auckland Bioengineering Institute, University of Auckland, New Zealand.

[12]Centre for Genomic Regulation and University Pompeu Fabra, Barcelona, Spain.

[13]Imperial College Healthcare NHS Trust, DuCane Road, London, United Kingdom.

[14]Ealing Hospital NHS Trust, Southall, Middlesex, United Kingdom.

[15]Pathology and Laboratory Medicine, University of Western Australia, Crawley, Western Australia, Australia.

[16]Department of Laboratory Medicine, Laboratory of Hematology, Radboud University Medical Center, Nijmegen, the Netherlands.

[17]Assistance Publique-Hopitaux de Paris, Institut National de la Santé et de la Recherche Médicale U1009, Villejuif, France.

[18]Biomedical Research Centre, Norwich Medical School, University of East Anglia, Norwich, United Kingdom.

[19]Center for Molecular and Vascular Biology, University of Leuven, Leuven, Belgium.

[20]Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[21]Department of Haematology, Hammersmith Campus, Imperial College Academic Health Sciences Centre, Imperial College London, London, United Kingdom.

[22]Department of Twin Research & Genetic Epidemiology, Genetics & Molecular Medicine Division, St Thomas' Hospital, King’s College, London, United Kingdom

[23]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom.

[24]Department of Oncology, Addenbrooke's Cambridge University Hospital NHS Trust, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[25]Cancer Research United Kingdom, Cambridge Institute, Cambridge Biomedical Campus, Cambridge, United Kingdom.

[26]School of Clinical Sciences, University of Bristol, United Kingdom.

[27]Genome Biology and Developmental Biology Units, European Molecular Biology Laboratory, Heidelberg, Germany.

[28]Wellcome Trust - Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom.

[29]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom.
* Correspondence to: ar506@cam.ac.uk (A.R.), mf471@cam.ac.uk (M.F.).
† These authors contributed equally to this work.
‡ A list of the members of the BRIDGE Bleeding and Platelet Disorders Consortium is available from Supplementary Materials.
[§] Sylvia Nürnberg is currently at Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.
[¶] Katrin Voss is currently at Loxxess Neuburg GmbH, Augsburger Straße 133, 86633 Neuburg an der Donau, Germany.
± Tiphaine Martin is currently at the Department of Twin Research & Genetic Epidemiology, Genetics & Molecular Medicine Division, St Thomas' Hospital, King's College, London, United Kingdom.

Abstract: Blood cells derive from hematopoietic stem cells through stepwise fating events. To characterize gene expression programs driving lineage choice we sequenced RNA from eight primary human hematopoietic progenitor populations representing the major myeloid commitment stages and the main lymphoid stage. We identify extensive cell-type specific expression changes: 6,711 genes and 10,724 transcripts, enriched in non-protein coding elements at early stages of differentiation. In addition, we discovered 7,881 novel splice junctions and 2,301 differentially used alternative splicing events, enriched in genes involved in regulatory processes. We demonstrate experimentally cell specific isoform usage, identifying NFIB as a regulator of megakaryocyte maturation − the platelet precursor. Our data highlight the complexity of fating events in closely related progenitor populations, the understanding of which is essential for the advancement of transplantation and regenerative medicine.

Main Text:

Introduction: Hematopoiesis has been extensively studied as a paradigm of stem cell biology and development (*1*). Hematopoietic stem cells (HSCs) and their progeny have been used to pioneer stem cell therapies for malignant and non-malignant hematological diseases (*2*) and the successful transplantation of genetically repaired HSCs is at the forefront of regenerative medicine in primary immune deficiency and severe combined immunodeficiency (*3*, *4*). HSCs reside in the bone marrow and can undergo asymmetric cell division (*5*), thereby generating an identical copy and a multipotent progenitor cell (MPP). MPPs have the ability to generate all hematopoietic cell types, but are incapable of indefinite self-renewal and engraftment (*6*, *7*). This process of expansion, differentiation and maturation culminates in the daily release of up to 10^{11} newly formed cells into the circulation, mainly red blood cells (RBCs) and platelets (*8*, *9*). The molecular mechanisms driving hematopoiesis have been classically understood as a cascade of gene expression programs propelled by transcription factors (TFs) (*10*) that direct lineage commitment and maturation by the coordinated regulation of gene transcription. Studies of hematological malignancies and model organisms (*1*) have identified many of the critical genes and mechanisms regulating hematopoietic development. Owing to species-specific differences, model organisms only contribute partially towards the detailed characterization of transcriptional cascades regulating human hematopoiesis (*11-13*).

Genome-wide transcriptional profiling of human hematopoietic progenitor populations has identified several transcriptional networks coordinating blood formation (*14*). However, gene expression datasets using whole-genome expression arrays only produce an incomplete assessment of the full repertoire of transcript isoforms that underpin the fating and expansion of progenitor cells (*14-16*). Alternative splicing is a widespread post-transcriptional process in eukaryotic organisms where multiple distinct transcripts are produced from a single gene (*17*). Analysis of RNA sequencing (RNA-seq) has shown that alternative splicing is used in up to 94% of human multi-exonic genes (*18*, *19*), often in a tissue and developmental stage-specific manner (*18*, *20*, *21*).

Alternative splicing has an important role in disease with 15% of disease-causing mutations located within splice sites and more than 20% of missense mutations lying within predicted splicing elements (*22*). Studies have also revealed that somatic mutations of splicing factor genes occur frequently in hematological cancers, including myelodysplasia and chronic lymphocytic leukemia (*23-25*). Thus, knowledge of cell-type specific alternative splicing and transcript isoform usage is required to interpret the consequences of genetic variation and to inform strategies for therapeutic intervention based on gene repair.

Results:

Deep transcriptomes of human hematopoietic progenitors

We used fluorescence-activated cell sorting of umbilical cord blood (CB) mononuclear cells to obtain highly purified populations of HSCs and five progenitor cells (MPP; CLP, common lymphoid progenitor; CMP, common myeloid progenitor; GMP, granulocyte monocyte progenitor; MEP, megakaryocyte erythrocyte precursor). In addition, erythroblasts (EBs) and megakaryocytes (MKs), the nucleated precursors of red blood cells (RBCs) and platelets, were obtained by in vitro differentiation of CB CD34+ cells (Figs. 1A, S1). For simplicity, we address all eight types of cells as progenitors.

We sequenced 25 poly-(A)$^+$ RNA samples, yielding a total of 2.4 x 10^9 uniquely aligned reads, ranging from 36 to 150 x 10^6 reads per sample (Table S1). We employed a Bayesian framework implemented in MMSEQ (*26*) to quantify gene and transcript expression by aligning reads to the transcriptome (Fig. 1B). Transcript usage ratio, the proportion of a gene's expression contributed by each of its transcripts, was also estimated (Fig. 1C). The latter provides an alternative to assessing differential transcript usage, which is less sensitive to depth of coverage and data normalization. To validate MMSEQ transcript expression estimates we purified 16 additional samples, representing 5 cell types, and performed quantitative RT-PCR (RT-qPCR) using the Fluidigm BioMark HD system for 36 transcript specific-assays. Linear regression between RNA-seq and RT-qPCR expression estimates indicated high reproducibility in biological replicates ($R^2 = 0.70$) (fig. S2, Table S2).

We confirmed the identity of each cell population by assessing the expression patterns of a set of well-characterized TF genes that are known to be essential for lineage commitment (Figs. 2A, S3) (*1*, *27*). For instance, *EBF1* expression peaks in CLPs, as expected from its role in B-cell development (*28*). Moreover the estimated gene expression of *GATA1* and *GATA2* reflects their switch in the differentiation of MEPs to EBs and MKs (*29*).

Classification of differential expression patterns during lineage commitment

To assess differential expression during hematopoietic lineage commitment at each branching point, such that all possible patterns of expression changes are considered, we used MMDIFF (*30*) to perform Bayesian polytomous model selection between the five possible modes of expression change involving three cell types (see **Materials and Methods** for further details and significance thresholds, Fig. 2B). This methodology identifies, for example, transcripts that are downregulated during the transition from CMP to GMP but retain similar expression between CMP and MEP.

Across all fating events we detected 6,711 genes, 10,724 transcripts and 7,017 transcript usage ratios with significant differences at least at one of the branching points (Fig. 2B). In total, we detected transcriptional changes per cell type in 22-33% of the 20,459 genes expressed across our dataset (defined as expression level ≥ 1 fragment per kilobase of transcript per million fragments mapped [FPKM] in at least two samples). Changes at the transcript level did not imply measurable differences at the gene level. The overlap between sets of differentially expressed transcripts (at the transcript and usage ratio levels) and of the genes they belong to was low, ranging from 0% to 35% (fig. S4). The extent of overlap did not increase when the threshold applied to ascertain expression (i.e. FPKM ≥ 1) was relaxed. Our analysis strategy highlights the advantages of using RNA-seq for assessing the richness of changes in expression at gene and transcript levels compared to probe-based technologies.

Of the 54,386 transcripts, expressed at an FPKM≥1 in at least two samples, 28,563 (52.5%) were protein-coding. The second and third most abundant classes were transcripts with retained introns (8,661, 15.9%) and processed transcripts without open reading frames (8,140, 15.0%). Assessment of the transcript biotypes of differentially expressed transcripts revealed that some modes of expression, at specific branching points, were significantly enriched for non-protein coding isoforms (Fig. 2C, Table S3). For example, during the HSC to MPP transition, transcripts upregulated in HSCs were enriched for non-protein-coding biotypes, such as lincRNAs (FDR = 0.043), whereas transcripts with similar expression in both cell types were enriched for protein-coding biotypes (FDR = 0.014). In contrast, differentially expressed transcripts at the terminal differentiation stage (MEP to EB or MK branching point) were enriched for transcripts with protein-coding biotypes (FDR < 0.016). These results suggest that a proportion of the regulation of lineage commitment, in the early stages of hematopoiesis, involves non-protein-coding elements and that lincRNAs may counteract differentiation programs, as observed in ES cells (*31*).

Cell-type specific genes and transcripts

Having evaluated expression patterns for genes differentially expressed between cell types at a branching point, we next focused on those genes and transcripts that were more highly expressed in one given cell population while displaying similar levels of relatively low expression in all other seven cell types (cf. **Materials and Methods**).

These cell-type specific genes or transcripts are likely to be important in conferring cellular identity (Fig. 2D, Tables S4, S5). Using conservative thresholds on the posterior probability and the extent of differential expression (cf. **Materials and Methods**), we identified between 6 (for MPP) and 631 (for EB) genes that were cell-type specific. We tested whether our cell-type specific gene sets were able to discriminate between cell populations in two microarray atlases of gene expression in human hematopoiesis (*14, 16*), achieving high concordance (fig S5 A-D). The number of cell-type specific transcripts ranged from between 19 for MPPs (belonging to 18 genes) and 1,807 for EBs (belonging to 1,141 genes). The low number of cell-type specific transcripts in MPPs is consistent with the small number of transcripts we identified as upregulated in this cell type (Fig. 2B). Thus MPP not only displayed a lower number of upregulated genes compared to HSCs, CMPs and CLPs, but also an overall lower number of cell-specific transcripts, suggesting a less distinct transcriptional identity of MPPs compared to other progenitor cells.

Consistent with our findings at branching events, cell-type specific gene and transcript sets show different patterns of enrichment of biotypes. Non protein-coding biotypes were over-represented in HSC-specific transcripts (FDR = 1.48×10^{-3}), while protein-coding transcripts were significantly enriched among transcripts specific in cells at terminally differentiated stages (EBs: FDR < 1.00×10^{-60}; MKs: FDR = 1.96×10^{-55}).

These cell-specific genes may play important roles in determining cell identity and proliferation of the different mature blood cells. We tested the hypothesis that these genes are directly implicated in mature cell differentiation and proliferation by overlapping these sets with genes harboring variants associated with RBC (*32*) and platelet (*33*) quantitative traits through genome wide association studies. Genes near loci associated with platelet count and volume were enriched in the MK-specific gene set (FDR=$1.7\text{x}10^{-8}$). In contrast, genes in loci associated with RBC count, volume and hemoglobin concentration were not enriched in the EB-specific gene set (FDR=0.42) or in any other cell specific set. This suggests that the regulation of platelet production is primarily intrinsic to MKs, while RBC production is regulated by mechanisms extrinsic to the erythroid lineage.

Our data add to the repertoire of genes and transcripts associated with cell identity in early and late stages of cell development in hematopoiesis, informing downstream examination of the role of transcriptional isoforms expressed in each cell population and their changes at each lineage commitment event.

Discovery and characterization of unannotated splice junctions

Owing to their low abundance and anatomical compartmentalization in the bone marrow, blood progenitor cells are systematically under-represented in existing transcript sequence databases. We analyzed an average of 137 million aligned reads per sample obtained across the 25 samples (Table S1, Fig. 1B) to explore the magnitude of unannotated splice junctions in human hematopoietic progenitors. We intersected splice junction calls from three splice-aware aligners (fig. S6) and required the splice junction to be observed in at least two samples. A total of 159,495 unique splice junctions were detected, of which 29,736 were not annotated in Ensembl v70. We categorized these unannotated splice junctions into four classes depending on whether their donor and acceptor sites were annotated within Ensembl (Fig. 3A). For 8,382 (28.2%) junctions both donor and acceptor sites were known but not the splicing pattern. 8,112 (27.3%) and 8,321 (28.0%) splice sites included unannotated splice donors or acceptors. Finally, 4,921 (16.5%) splice events had both donor and acceptor sites unannotated. The frequency of the four different categories of splicing events did not differ between the eight cell populations (Fig. 3A).

To characterize the 29,736 putative unannotated splice junctions, we investigated their splice site probability scores, degree of conservation and coding potential. Splice site probability scores (*34*) for unannotated splice sites were similar in known and unannotated donor sites (>0.90, fig. S7). We observed that conservation scores for exonic regions with unannotated splice sites were higher (mean 0.28) than for intronic regions in the same splice sites (mean 0.20; $P<2.2\text{x}10^{-16}$; Wilcoxon rank sum test, fig. S8). Finally, protein-coding potential of the unannotated exons was assessed with the frequency of stop codons, in all three reading frames for both directions, within a 100 base pairs (bp) window around the splice sites. A similar distribution of coding potential was identified in unannotated and annotated exons, with both containing an average of 1.2 fewer stop codons than the equivalent intronic regions (fig. S9).

To identify novel splice junctions, we compared the unannotated splice junctions to splice junctions identified in the UCSC ESTs/mRNA dataset and in poly-$(A)^+$ RNA-seq dataset from 16 human tissues in Illumina BodyMap 2.0 (*35*). In total, 73.5% of our unannotated events were detected in these external datasets, with 23.0% detected in UCSC ESTs/mRNA data and 72.0% in our re-analysis of the BodyMap 2.0 dataset (fig. S10). The remaining 7,881

(26.5%) splice junctions were specific to our dataset (hereafter called novel). Analysis of novel splice junctions revealed a higher proportion of the non-canonical splicing motif GC-AG (2.2% and 7.3% in unannotated and novel, respectively) compared to annotated sites (0.9%, fig. S11). While both the GT-AG and GC-AG splice sites are processed by the canonical U2-type spliceosome, GC-AG splice sites tend to be alternatively spliced (*36*).

We calculated Shannon's entropy (*37*) for the three classes of splice junctions: annotated, unannotated and novel (Fig 3B). A lower entropy distribution in the novel splice junction set indicates that these tend to be population-specific events, when compared to the unannotated splice junctions present in BodyMap 2.0 data, or all annotated junctions ($P<2.2x10^{-16}$; Wilcoxon rank sum test, fig. S12). Enrichment analysis of genes containing novel splice junctions highlighted GO terms related to cell cycle, DNA metabolism and RNA processing ($FDR < 5.0x10^{-17}$, Table S8), suggesting that these novel splice junctions may alter the function of genes involved in critical cellular processes.

Our results suggest that a number of the unannotated and novel splice junctions are indeed functional and used in a cell-type specific manner based on their splice probability, conservation score, coding potential and entropy.

Validation of novel splice junctions

We used PCR to independently validate 23 of the novel splice junctions. The PCR assays (Table S9) performed on 5 samples showed > 90% concordance rate (105/115 reactions, figs. S13-S16, Table S10). We performed additional sequencing of poly-$(A)^+$ transcripts from sample MK_3 using the PacBio RS sequencing platform, which enables sequencing of full-length transcripts and overcomes the limitations of transcriptome assembly on the basis of short Illumina reads (*38*). PacBio sequencing yielded 67,110 reads identified as full-length molecules (originating from individual RNA transcripts) ranging between 322 and 13,170 bp in length (median 2,272 bp). These were further combined into 35,663 consensus sequence clusters – transcript structures (cf. **Material and Methods**). Two novel splice junctions validated by PCR in MKs were also observed within the PacBio dataset (fig. S15). Using these data we investigated the transcriptional context of the novel splice junctions in MKs. Visual inspection of the PacBio alignments indicated that a number novel splice junctions are part of full-length transcripts, including a previously unobserved intergenic locus on chromosome 12 and an antisense transcript within an annotated protein encoding region in the *GNG12* locus (fig. S15).

Of the 94,423 splice junctions with 10 or more Illumina reads in MK_3, 54% were supported by PacBio data. In contrast, 7% (66/956) of novel and 11% (773/7,234) of unannotated splice junctions identified in MK_3 were recapitulated in the PacBio dataset. We used the annotated splice junctions to estimate the probability of detection by PacBio as a function of read depth and transcript length. The observed validation rates of unannotated and novel junctions, after accounting for read depth, would be consistent with the majority of these junctions originating from transcripts less than 300 bp in length (fig. S17 and (*39*)). Notwithstanding PacBio's lower depth of sequencing and other unaccounted technical aspects, this analysis provides support to the idea that a large fraction of novel splicing events involve very short transcripts not captured by PacBio.

Differential usage of alternative splice junctions

To investigate the prevalence of cell-type specific alternative splicing, we identified 42,001 splice junction sets where two or more splice junctions shared either the donor or the acceptor sites. Of these, we focused on the 20,924 (49.8%) junctions that contained only two splicing alternatives and were detected in at least two biological replicates. To determine if an alternative splice site displays differential splicing usage (DSU), that is, that the relative contribution of splicing alternatives (usage proportion) differs between a given cell-type and the average proportion across all other cell types, we fitted a beta binomial model and established statistical significance using a likelihood ratio test. The beta binomial model accounts for the overdispersion - beyond the expected binomial variance - present in the data. It is an extension to the binomial model (i.e. logistic regression) and is akin to using a negative binomial distribution to model overdispersed counts data. This analysis identified 2,301 DSU sets ($FDR<0.1$). The number of DSU events ranged between 4 for HSC and 1,034 for CLP (Table S11, figs. S14, S18, S19). The DSU set was enriched with novel splice junctions compared to all junctions ($P= 4.39x10^{-7}$, Chi-square test).

To better characterize the biological relevance of cell-type specific DSU, we classified splicing events according to their transcriptional consequences: 73.4% lead to exon skipping events, 8.3% of junctions have an alternative 3' acceptor, and 6.2% have an alternative 5' donor. 12.1% of events could not be annotated using the reference transcriptome (fig. S20). Although unannotated, the length distribution of this fraction suggests that the majority is composed of exon skipping events (fig. S20).

In the alternative spliced regions displaying DSU, 26.1% contain a premature stop codon and 38.5% contained at least one predicted protein structure or domain, therefore resulting in gain or loss of protein functions. No one type of domain was significantly overrepresented in the DSU set. Of the alternative acceptor sites displaying DSU, 39% (84/216) resulted from a 3 bp shift in the alternative acceptor sites (fig. S21, left panel) displaying a NAGNAG motif (*40*) (fig. S21, right panel). This motif maintains the translation frame and may introduce a single amino-acid (aa) insertion or a substitution (fig. S22). The 2,301 DSU events could be assigned to 1,704 genes. GO enrichment analysis of these genes indicates that these genes may be directly involved in the regulation of transcription and splicing (Table S12).

We validated 11 DSU events, four novel and seven known events, using PCR (figs. S14, S15, S23). Densitometry estimates of the percentage-spliced-in (PSI) of these PCRs correlates with the PSI estimated from the RNA-seq data (N=26, R^2=0.78, fig. S24). In addition we validated a novel DSU in *NFIB* (see below and, fig. 4a) and a DSU event in *GFI1B* (*39*). The DSU event in *GFI1B*, in CMPs, EBs and MKs results in an alternatively spliced-out exon 4 that encodes for two Zn finger domains critical for megakaryopoiesis (*39*). Overall, the DSU analysis confirms alternative splicing as an additional key mechanism through which fundamental processes during hematopoiesis are regulated.

RNA-binding motif enrichment in DSU

Alternative splicing is regulated by trans-acting splicing factors that recognize cis-acting sequences in exons or introns, to promote or suppress the assembly of the spliceosome at the adjacent splice site. We therefore investigated the molecular regulation of cell-specific alternative splicing by examining the sequences around alternatively spliced exons. We used 102 recently described RNA-binding motifs of 80 human RNA-binding proteins (*41*) to identify sets of motifs significantly enriched or depleted in the regions surrounding DSU junctions (Table S13). Of the 80 RNA-binding proteins with known binding motifs, 59 were expressed in our data with FPKM>1 and displayed variable cell-type specificity (fig. S25). RNA-binding motif enrichment analysis was performed on cassette exons and proximal intronic regions. The patterns of enrichment and depletion, in addition to the identity of the motifs, varied widely across cell types (Fig. 3C).

The proteins BRUNOL, SRSF, TIA1 and the HNRNP (heterogeneous ribonucleoprotein) family of proteins are known to regulate tissue-specific splicing (*42*). The patterns of enrichment and depletion in our dataset for these proteins suggest their role in regulating tissue-specific splicing also extends to hematopoietic cells (Fig. 3C). For example, we identified that the motifs of the HNRNP protein family, which typically bind to exonic splicing silencers (*43*), were enriched in exonic regions of MPPs and MEPs that are spliced out.

Novel isoform of NFIB regulates megakaryopoiesis

To investigate the impact of different transcript isoforms in a biological system we focused our attention on the role of two TFs in megakaryopoiesis: *NFIB*, described below, and *GFI1B* (*39*), as an example of how our analysis informs the interpretation of patient sequencing data. *NFIB* was identified at the MEP/EB/MK branching point (fig. S26), containing a novel MK-specific DSU event (FDR < 0.05). The role of *NFIB* has been extensively studied in lung maturation, the nervous system (*44*) and epithelial stem cell development (*45*). The *NFI* family of TFs, constituted by four members (A, B, C and X), has previously been implicated in regulating hematopoiesis: with *Nfix* identified as functional in murine HSCs and progenitors (*46*), and *NFIA* implicated in human erythropoiesis (*47*). *NFIC* has been observed as being differentially expressed between MKs of fetal and postnatal origin (*48*). In addition, *NFIB* has been identified as one of the TFs down-regulated in the HSC to MPP transition (*49*). However, its role in the later stages of hematopoiesis has remained unexplored.

By examining genomic alignments we identified a novel *NFIB* transcript (chr9:14,179,779-14,214,332bp) and annotated the position of the transcription start site (TSS) in the novel first exon. The isoform that results from this novel transcript was primarily expressed in HSCs and MKs, and was only present in white blood cells in the BodyMap 2.0 dataset, while the canonical isoform is widely expressed across other BodyMap 2.0 tissues. The novel TSS lies in a region of open chromatin in primary MKs (*50*) that is occupied by the TFs MEIS1 (this study), FLI1, GATA1, SCL/TAL1, but not GATA2 or RUNX1 (*51*). The TSS is also marked by the promoter mark H3K4me3 in MKs (fig. S27). We validated the novel TSS by 5' race RT-PCR and observed multiple PacBio reads supporting it (Fig. 4A). Western blotting (WB) confirmed the presence of the protein encoded by the novel short isoform, NFIB-

S, as the major isoform in MKs (Figs. 4B, S28) while the longer isoform encoded by the canonically spliced transcript could not be detected.

NFIB is known to bind DNA preferentially as a homodimer or a heterodimer in combination with other NFI family members(*52*). Since NFIB-S lacks the DNA binding and protein interaction domains (*53*), we investigated its ability to interact with NFIC in MKs, given its previously hypothesized role in definitive postnatal megakaryopoiesis (*48*). Co-transfection experiments followed by immunoprecipitation showed that the novel isoform, NFIB-S, lacked the ability to interact with NFIC (Fig. 4C). To determine the role of both NFIB and NFIC during megakaryopoiesis, we induced peripheral blood CD34+ cells to differentiate towards MKs and infected them with pools of shRNA lentiviruses targeting *NFIB*, *NFIC*, or a non-silencing control. Knockdown of either gene resulted in a marked reduction in differentiation towards MKs as assessed by flow cytometry (Fig. 4D) and confirmed by morphological analysis (fig. S29). This indicated that both NFIB-S and NFIC have an essential role in megakaryopoiesis despite the absence of a DNA binding domain in NFIB-S. Overexpression of both *NFIB*-S and *NFIC* in CD34+ cells increased cell maturation (Fig. 4E, P = 0.001 and P = 0.014, respectively), measured as double positivity for the MK maturation markers CD41a (ITGA2B) and CD42b (GP1BA) (*54*). In contrast, overexpression of the canonical isoform, *NFIB-L*, had no effect (Fig. 4E). These experiments indicate that both NFIC and the novel isoform of NFIB-S, identified in our analysis, play a critical interlinked role in the formation of MKs.

Discussion:

Current knowledge of gene expression and function in hematopoiesis is mainly based on observations at gene level. However, it is clearly the transcript, rather than the gene, to which biological function should be ascribed, either as protein-coding or as non-protein-coding RNA.

Here, RNA-seq of HSCs and seven progenitor populations has enabled the identification, quantification and differential expression analysis of cell-type specific transcript isoforms, novel and unannotated splice junctions and alternative splicing events at a genome-wide level. Analysis of lineage commitment events revealed a wealth of previously undetectable transcript switching and of shifts altering isoform usage ratio, without appreciable changes at gene level, providing evidence of additional layers of regulation in cell fating.

Generating an atlas of splicing events allowed us to explore the diversity and mechanisms behind alternative splicing in human hematopoiesis as well as to contribute further to the human genome functional annotation by reporting 7,881 novel splice junctions, specific to these rare cell populations.

To demonstrate the importance of specific isoforms in driving lineage fating events we investigated the role of a transcription factor highlighted by the polytomous analysis. We envisage that integration of this Blueprint RNA-seq dataset with the deep catalogues generated by Blueprint and other epigenome consortia, will aid the annotation of the functional genomic landscape of the hematopoietic system. This is essential in the continued effort to interpret the functional consequences of mutations in patients with rare hematological disorders and support the next enhancements of personalized treatments for patients with hematological malignancies.

Materials and Methods:

Progenitor cell purification

Cord blood (CB) was collected after informed consent (ethical approval REC 12/EE/0040) and the mononuclear cells extracted. CD34+ cells were isolated using the EasySep® progenitor cell enrichment kit with platelet depletion (STEMCELL Technologies, Vancouver, Canada), stained using a panel of antibodies and flow sorted to purify HSC, MPP, CLP, CMP, MEP and GMP cells that were lysed in TRIZOL reagent (Life Technologies, Carlsbad, CA, USA).

Cell culture and purification

EBs and MKs were cultured from CD34+ cells isolated from CB mononuclear cells with the human CD34 isolation kit (Miltenyi Biotec, Bergisch Gladbach, Germany). For EBs, CD34+ cells were cultured with erythropoietin, SCF and IL3 for 14 days. For MKs CD34+ cells were cultured for 10 days in thrombopoietin and IL1β. Both populations were immuno-selected to > 95% purity before lysis.

RNA-seq library preparation and sequencing

RNA was extracted from TRIZOL preparations. 100 pg of RNA was used to generate poly-(A)$^+$ RNA libraries with the SMARTer Ultra Low RNA and Advantage 2 PCR kits (Clontech, Mountain View, CA, USA). Samples were indexed with NEXTflex adapters (Bioo-scientific, Austin, TX) and 100 base-pair paired-end sequencing was performed on Illumina HiSeq 2000 instruments with TruSeq reagents (Illumina, San Diego, CA, USA).

Quality control, trimming, alignment and expression analysis

RNA-seq libraries were initially subjected to a quality control step, where outliers were identified and discarded from further analysis on the basis of the duplication rates and gene coverage. Paired-end reads of the 25 independent samples were trimmed for both PCR and sequencing adapters with Trim Galore (*55*). Trimmed reads were aligned to the Ensembl v70 human transcriptome using Bowtie (*56*). Quantification of gene and transcript expression was performed with MMSEQ (*26*)

Differential expression analysis through polytomous model classification

Presence of significant differential expression was determined with MMDIFF (*30*) at three different levels: gene, transcript and isoform usage (generically called features).

Polytomous classification was carried out by first performing two-model comparisons to calculate Bayes factors, *B(0,m)*, between a common baseline model and models representing the expression patterns of interest for a given feature. Under the baseline model, *0*, the feature's mean expression level is the same in all cell types and under the alternative model, *m*, the mean expression level is allowed to differ according to the desired pattern (e.g. CMP=GMP≠MEP). Bayes' theorem was used to compute the posterior probability that the true model γ is equal to *m* under the assumption that the alternative models are exhaustive: $P(\gamma = m|x) = B(0,m)\times P(\gamma = m) / \sum B(0,m')\times P(\gamma =m')$, where *x* denotes the MMSEQ estimates for that feature.

For the transition from HSCs to MPPs, we used a two-model comparison, where we used a prior probability that the baseline model was true of 0.9. This can be interpreted as a prior belief that 10% of features are differentially expressed. Features with a posterior probability for the alternative model above 0.5 (equivalent to a Bayes factor threshold of 9, representing strong evidence for the alternative model) and an FPKM > 1 in at least two of the samples involved, were considered differentially expressed.

At each cell-fating point involving three cell types, we studied all patterns of expression amongst the progenitor cell and its immediate progeny. We classified feature expression patterns according to five models. The simplest model assumes that the mean expression level is the same across cell types. The most complex model assumes that the mean expression level is different for each cell type. The remaining three models assume that two of the three cell types have the same mean expression level. We specified a prior probability of 80% for the simplest model and distributed the remaining probability evenly across the four alternative models. The model with the highest posterior probability was selected.

Gene set enrichment analysis

Gene and transcript sets derived from the polytomous and the cell-type specific expression analyses of the RNA-seq data were tested for gene set enrichment with the goseq R/Bioconductor package version 1.14.0 (*57*), which accounts for the relationship between power of detection and transcript length. All P values were corrected for multiple testing using the Benjamini-Hochberg method (*58*).

Selection of Cell-type specific genes and transcripts

We selected cell-type specific genes and transcripts by performing a 9-model polytomous comparison. The simplest model assumes that the mean expression level is the same across cell types. Each of the remaining 8 models assumes that the expression is the same across all cell types except for one of the progenitors. We specified a prior probability of 0.5 for the simplest model and distributed the remaining probability evenly across the eight alternative models. Genes and transcripts were required to have a posterior probability greater than 0.5 and a fold change in expression greater than 1 in order to be declared cell-type specific. To compare cell-type specific gene expression estimates between our RNA-seq data and publicly available microarray datasets, we retrieved probe annotations for the Illumina (*16*) and Affymetrix (*14*) platforms from Ensembl v70.

Splice junction analysis

Identification of splice junctions for each sample was based on the alignment of the trimmed reads to the human genome (GRCh37) with three different aligners: GSNAP (*59*), STAR (*60*) and GEM (*61*). Splice junctions were considered for further analysis if supported by all three aligners and by at least 10 reads in at least two samples, where reads covered a minimum of 10-bp at both ends of the splice junction. We defined a splice junction as unannotated if not present in Ensembl v70. These were further compared to the EST/mRNA data from UCSC and the Illumina BodyMap 2.0 dataset (*35*) to identify novel splice junctions.

Splice site probability scores were extracted from the GSNAP output. PhastCons conservation scores were used to plot evolutionary conservation in the 100-bp surrounding each splice junction. Coding potential was estimated by summing the number of possible stop codons in exonic and intronic regions, in all reading frames and in 100-bp flanking the unannotated splice site. Shannon's entropy (*37*) was calculated on the basis of the read coverage of the splice junctions.

DSU was identified with a beta binomial model. The characterization of the protein domains in cassette exons with DSU was performed with InterProscan 5 (*62*) to search for domains predicted by Pfam.

Validation of transcript isoform expression and splicing events

To validate the quantification of transcript levels determined by analysis of the RNA-seq data with MMSEQ, we performed RT-PCR assays with 40 transcript-specific assays and five positive control assays in multiple cell subsets. Quantitation of each transcript was performed in multiple progenitor cell subsets with the BioMark HD system (Fluidigm, San Francisco, CA). After requiring call quality scores > 0.9 (Fluidigm Real-Time PCR Analysis software, http://www.fluidigm.com/software.html), 36 transcripts were analyzed. For each probe, □□Cq values were calculated with the B2M transcript as control and the average □Cq for MKs. Linear regression was then performed between □□Cq values and the corresponding MMSEQ estimates relative to mean MK.

To validate progenitor-specific novel splice junctions and exon-skipping events we designed PCR primers to amplify 30 junctions identified by RNA-seq. PCR was performed on pools of the RNA-seq libraries. PCR products were run on agarose gel, imaged and densitometry performed.

PacBio libraries were generated (Pacific Biosciences, Menlo Park, CA) from cDNA obtained by reverse transcription of 10 ng of MK_3 total RNA and sequenced in five SMRTcells on the PacBio RSII. SMRTpipe and ICE were used to filter reads to generate consensus sequence clusters that were mapped to the reference human genome (GRCh37) using GMAP.

Enrichment analysis of RNA-binding motifs around cassette exons

Motif enrichment analysis of 102 RNA binding motifs (*41*) was performed on DSU cassette exons (FDR < 0.05 and usage proportion change > 0.05) over three genomic regions: upstream intronic (300-bp), exonic and downstream intronic (300-bp). Enrichment and depletion of RNA binding motifs was determined with cumulative hypergeometric testing and P values were corrected for multiple testing.

Cloning, shRNA, lentivirus production and transduction

TRC shRNA lentivirus targeting *NFIB* and *NFIC* and a non-silencing control were purchased from Thermo Scientific Open Biosystems (Little Chalfont, UK). NFIB full length and NFIC cDNA were cloned into pWPI TAP tagged vector. Packaging was performed in 293T cells and viral stocks were titrated and quantified using qPCR and for pWPI using qPCR and GFP FACS. CD34+ cells were purified from NHS Blood and Transplant apheresis filters, as above (Miltenyi), and then infected with lentiviral particles in the presence of polybrene, in media supplemented with thrombopoietin and IL1b. On day 2, media was replaced and cells cultured to MKs. At day 10, MKs were counted and assessed by morphology and flow cytometry for maturation.

Transfections, immunoprecipitations and Western blots

To detect protein-protein interactions, NFI proteins were expressed by co-transfection in 293T cells and immunoprecipitated with an anti-flag antibody. Western blots were probed with NFIB, NFIC, β-Tubulin.

Data presentation

River plots were generated with the ggplot2 R package (version 0.9.3.1), which was obtained from Bioconductor (http://www.bioconductor.org/). Heatmaps were generated with both the gplots (version 2.13.0) and pheatmap (version 0.7.7) R packages, which were both obtained from CRAN (http://cran.r-project.org/). For sequence logos of the splice site motifs we used seqLogo R package (version 1.30.0) available from Bioconductor. The IGV genome browser (version 2.3.34) was used for visualization (http://www.broadinstitute.org/igv/).

For further details please refer to the supplementary material.

References and Notes:

1. S. H. Orkin, L. I. Zon, Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631-644 (2008). doi: 10.1016/j.cell.2008.01.025.
2. J. Hoggatt, J. M. Speth, L. M. Pelus, Sowing the seeds of a fruitful harvest: Hematopoietic stem cell mobilization. *Stem cells (Dayton, Ohio),* (2013). doi: 10.1002/stem.1574.
3. Slukvin, II, Hematopoietic specification from human pluripotent stem cells: current advances and challenges toward de novo generation of hematopoietic stem cells. *Blood* **122**, 4035-4046 (2013). doi: 10.1182/blood-2013-07-474825.
4. M. P. Lambert, S. K. Sullivan, R. Fuentes, D. L. French, M. Poncz, Challenges and promises for the development of donor-independent platelet transfusions. *Blood* **121**, 3319-3324 (2013). doi: 10.1182/blood-2012-09-455428.
5. B. Giebel, I. Bruns, Self-renewal versus differentiation in hematopoietic stem and progenitor cells: a focus on asymmetric cell divisions. *Current stem cell research & therapy* **3**, 9-16 (2008). doi.

6. R. Majeti, C. Y. Park, I. L. Weissman, Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell stem cell* **1**, 635-645 (2007). doi: 10.1016/j.stem.2007.10.001.
7. C. M. Baum, I. L. Weissman, A. S. Tsukamoto, A. M. Buckle, B. Peault, Isolation of a candidate human hematopoietic stem-cell population. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 2804-2808 (1992). doi.
8. S. H. Orkin, Diversification of haematopoietic stem cells to specific lineages. *Nature reviews. Genetics* **1**, 57-64 (2000). doi: 10.1038/35049577.
9. S. Doulatov, F. Notta, E. Laurenti, J. E. Dick, Hematopoiesis: a human perspective. *Cell stem cell* **10**, 120-136 (2012). doi: 10.1016/j.stem.2012.01.006.
10. H. Iwasaki *et al.*, The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes & development* **20**, 3010-3021 (2006). doi: 10.1101/gad.1493506.
11. J. Mestas, C. C. Hughes, Of mice and not men: differences between mouse and human immunology. *Journal of immunology (Baltimore, Md. : 1950)* **172**, 2731-2738 (2004). doi.
12. D. G. Tenen, R. Hromas, J. D. Licht, D. E. Zhang, Transcription factors, normal myeloid development, and leukemia. *Blood* **90**, 489-519 (1997). doi.
13. R. E. Dickinson *et al.*, Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood* **118**, 2656-2658 (2011). doi: 10.1182/blood-2011-06-360313.
14. N. Novershtern *et al.*, Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296-309 (2011). doi: 10.1016/j.cell.2011.01.004.
15. N. A. Watkins *et al.*, A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* **113**, e1-9 (2009). doi: 10.1182/blood-2008-06-162958.
16. E. Laurenti *et al.*, The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature immunology* **14**, 756-763 (2013). doi: 10.1038/ni.2615.
17. T. W. Nilsen, B. R. Graveley, Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457-463 (2010). doi: 10.1038/nature08909.
18. E. T. Wang *et al.*, Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008). doi: 10.1038/nature07509.
19. Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**, 1413-1415 (2008). doi: 10.1038/ng.259.
20. N. L. Barbosa-Morais *et al.*, The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)* **338**, 1587-1593 (2012). doi: 10.1126/science.1230612.
21. J. Merkin, C. Russell, P. Chen, C. B. Burge, Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science (New York, N.Y.)* **338**, 1593-1599 (2012). doi: 10.1126/science.1228186.
22. R. K. Singh, T. A. Cooper, Pre-mRNA splicing in disease and therapeutics. *Trends in molecular medicine* **18**, 472-482 (2012). doi: 10.1016/j.molmed.2012.06.006.
23. K. Yoshida *et al.*, Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69 (2011). doi: 10.1038/nature10496.
24. M. Cazzola, M. G. Della Porta, L. Malcovati, The genetic basis of myelodysplasia and its clinical relevance. *Blood* **122**, 4021-4034 (2013). doi: 10.1182/blood-2013-09-381665.
25. Y. Wan, C. J. Wu, SF3B1 mutations in chronic lymphocytic leukemia. *Blood* **121**, 4627-4634 (2013). doi: 10.1182/blood-2013-02-427641.
26. E. Turro *et al.*, Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome biology* **12**, R13 (2011). doi: 10.1186/gb-2011-12-2-r13.
27. A. C. Wilkinson, B. Gottgens, Transcriptional regulation of haematopoietic stem cells. *Advances in experimental medicine and biology* **786**, 187-212 (2013). doi: 10.1007/978-94-007-6621-1_11.
28. D. Bryder, M. Sigvardsson, Shaping up a lineage--lessons from B lymphopoesis. *Current opinion in immunology* **22**, 148-153 (2010). doi: 10.1016/j.coi.2010.02.001.
29. H. Kaneko, R. Shimizu, M. Yamamoto, GATA factor switching during erythroid differentiation. *Current opinion in hematology* **17**, 163-168 (2010). doi: 10.1097/MOH.0b013e32833800b8.
30. E. Turro, W. J. Astle, S. Tavare, Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics (Oxford, England)* **30**, 180-188 (2014). doi: 10.1093/bioinformatics/btt624.
31. M. Guttman *et al.*, lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011). doi: 10.1038/nature10398.

32. P. van der Harst *et al.*, Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369-375 (2012). doi: 10.1038/nature11677.
33. C. Gieger *et al.*, New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201-208 (2011). doi: 10.1038/nature10659.
34. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology* **11**, 377-394 (2004). doi: 10.1089/1066527041410418.
35. C. M. Farrell *et al.*, Current status and new features of the Consensus Coding Sequence database. *Nucleic acids research* **42**, D865-872 (2014). doi: 10.1093/nar/gkt1059.
36. T. A. Thanaraj, F. Clark, Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic acids research* **29**, 2581-2593 (2001). doi.
37. J. Schug *et al.*, Promoter features related to tissue specificity as measured by Shannon entropy. *Genome biology* **6**, R33 (2005). doi: 10.1186/gb-2005-6-4-r33.
38. T. Steijger *et al.*, Assessment of transcript reconstruction methods for RNA-seq. *Nature methods* **10**, 1177-1184 (2013). doi: 10.1038/nmeth.2714.
39. See supplemental material.
40. M. Hiller *et al.*, Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature genetics* **36**, 1255-1257 (2004). doi: 10.1038/ng1469.
41. D. Ray *et al.*, A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177 (2013). doi: 10.1038/nature12311.
42. Y. Barash *et al.*, Deciphering the splicing code. *Nature* **465**, 53-59 (2010). doi: 10.1038/nature09000.
43. D. L. Black, Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**, 291-336 (2003). doi: 10.1146/annurev.biochem.72.121801.161720.
44. G. Steele-Perkins *et al.*, The transcription factor gene Nfib is essential for both lung maturation and brain development. *Molecular and cellular biology* **25**, 685-698 (2005). doi: 10.1128/mcb.25.2.685-698.2005.
45. C. Y. Chang *et al.*, NFIB is a governor of epithelial-melanocyte stem cell behaviour in a shared niche. *Nature* **495**, 98-102 (2013). doi: 10.1038/nature11847.
46. P. Holmfeldt *et al.*, Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122**, 2987-2996 (2013). doi: 10.1182/blood-2013-04-493973.
47. L. M. Starnes *et al.*, A transcriptome-wide approach reveals the key contribution of NFI-A in promoting erythroid differentiation of human CD34(+) progenitors and CML cells. *Leukemia* **24**, 1220-1223 (2010). doi: 10.1038/leu.2010.78.
48. O. Bluteau *et al.*, Developmental changes in human megakaryopoiesis. *Journal of thrombosis and haemostasis : JTH* **11**, 1730-1741 (2013). doi: 10.1111/jth.12326.
49. F. Notta *et al.*, Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science (New York, N.Y.)* **333**, 218-221 (2011). doi: 10.1126/science.1201219.
50. D. S. Paul *et al.*, Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome research* **23**, 1130-1141 (2013). doi: 10.1101/gr.155127.113.
51. M. R. Tijssen *et al.*, Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Developmental cell* **20**, 597-609 (2011). doi: 10.1016/j.devcel.2011.04.008.
52. D. Apt, T. Chong, Y. Liu, H. U. Bernard, Nuclear factor I and epithelial cell-specific transcription of human papillomavirus type 16. *Journal of virology* **67**, 4455-4463 (1993). doi.
53. J. Dekker, J. A. van Oosterhout, P. C. van der Vliet, Two regions within the DNA binding domain of nuclear factor I interact with DNA and stimulate adenovirus DNA replication independently. *Molecular and cellular biology* **16**, 4073-4080 (1996). doi.
54. S. Poirault-Chassac *et al.*, Notch/Delta4 signaling inhibits human megakaryocytic terminal differentiation. *Blood* **116**, 5670-5678 (2010). doi: 10.1182/blood-2010-05-285957.
55. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
56. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009). doi: 10.1186/gb-2009-10-3-r25.
57. M. D. Young, M. J. Wakefield, G. K. Smyth, A. Oshlack, Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology* **11**, R14 (2010). doi: 10.1186/gb-2010-11-2-r14.
58. Y. Hochberg, Y. Benjamini, More powerful procedures for multiple significance testing. *Statistics in medicine* **9**, 811-818 (1990). doi.

59. T. D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)* **26**, 873-881 (2010). doi: 10.1093/bioinformatics/btq057.
60. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21 (2013). doi: 10.1093/bioinformatics/bts635.
61. S. Marco-Sola, M. Sammeth, R. Guigo, P. Ribeca, The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature methods* **9**, 1185-1188 (2012). doi: 10.1038/nmeth.2221.
62. P. Jones *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236-1240 (2014). doi: 10.1093/bioinformatics/btu031.
63. D. Karolchik *et al.*, The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**, D764-770 (2014). doi: 10.1093/nar/gkt1168.
64. T. D. Wu, C. K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)* **21**, 1859-1875 (2005). doi: 10.1093/bioinformatics/bti310.
65. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050 (2005). doi: 10.1101/gr.3715005.
66. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**, 1096-1098 (2013). doi: 10.1038/nmeth.2639.
67. J. C. Castle *et al.*, Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature genetics* **40**, 1416-1425 (2008). doi: 10.1038/ng.264.
68. S. T. Nurnberg *et al.*, A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood* **120**, 4859-4868 (2012). doi: 10.1182/blood-2012-01-401893.
69. A. Diaz, A. Nellore, J. S. Song, CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome biology* **13**, R98 (2012). doi: 10.1186/gb-2012-13-10-r98.
70. S. Saleque, S. Cameron, S. H. Orkin, The zinc-finger proto-oncogene Gfi-1b is essential for development of the erythroid and megakaryocytic lineages. *Genes & development* **16**, 301-306 (2002). doi: 10.1101/gad.959102.
71. L. Vassen *et al.*, Growth factor independent 1b (Gfi1b) and a new splice variant of Gfi1b are highly expressed in patients with acute and chronic leukemia. *International journal of hematology* **89**, 422-430 (2009). doi: 10.1007/s12185-009-0286-5.
72. M. Osawa *et al.*, Erythroid expansion mediated by the Gfi-1B zinc finger protein: role in normal hematopoiesis. *Blood* **100**, 2769-2777 (2002). doi: 10.1182/blood-2002-01-0182.
73. B. Laurent *et al.*, A short Gfi-1B isoform controls erythroid differentiation by recruiting the LSD1-CoREST complex through the dimethylation of its SNAG domain. *Journal of cell science* **125**, 993-1002 (2012). doi: 10.1242/jcs.095877.
74. D. Monteferrario *et al.*, A dominant-negative GFI1B mutation in the gray platelet syndrome. *The New England journal of medicine* **370**, 245-253 (2014). doi: 10.1056/NEJMoa1308130.
75. W. S. Stevenson *et al.*, GFI1B mutation causes a bleeding disorder with abnormal platelet function. *Journal of thrombosis and haemostasis : JTH* **11**, 2039-2047 (2013). doi: 10.1111/jth.12368.
76. S. A. Wolfe, R. A. Grant, M. Elrod-Erickson, C. O. Pabo, Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure (London, England : 1993)* **9**, 717-723 (2001). doi.
77. S. Kohler *et al.*, The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research* **42**, D966-974 (2014). doi: 10.1093/nar/gkt1026.
78. G. R. Abecasis *et al.*, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010). doi: 10.1038/nature09534.
79. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589-595 (2010). doi: 10.1093/bioinformatics/btp698.
80. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011). doi: 10.1038/ng.806.
81. G. Jun *et al.*, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics* **91**, 839-848 (2012). doi: 10.1016/j.ajhg.2012.09.004.
82. M. Westerfield, *The zebrafish book : a guide for the laboratory use of zebrafish (Danio rerio)*. (M. Westerfield, Eugene, OR, ed. 4, 2000).
83. R. Knight *et al.*, PyCogent: a toolkit for making sense from sequence. *Genome biology* **8**, R171 (2007). doi: 10.1186/gb-2007-8-8-r171.

84. A. Loytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science (New York, N.Y.)* **320**, 1632-1635 (2008). doi: 10.1126/science.1158395.
85. T. Massingham, N. Goldman, Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753-1762 (2005). doi: 10.1534/genetics.104.032144.
86. B. Sipos, T. Massingham, G. E. Jordan, N. Goldman, PhyloSim - Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC bioinformatics* **12**, 104 (2011). doi: 10.1186/1471-2105-12-104.
87. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007). doi: 10.1093/molbev/msm088.
88. C. UniProt, Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research* **41**, D43-47 (2013). doi: 10.1093/nar/gks1068.

Acknowledgements:

The authors wish to acknowledge the following funding bodies:
The work described in this manuscript was primarily supported by the European Commission Seventh Framework Program through the BLUEPRINT grant with code HEALTH-F5-2011-282510 (DH, FB, GC, JHAM, KD, LC, MF, SC, SF and SPG). Research in the Ouwehand laboratory is further supported by program grants from the National Institute for Health Research (NIHR, http://www.nihr.ac.uk; to AA, MK, PP, SBGJ, SN, and WHO); and the British Heart Foundation under numbers RP-PG-0310-1002 and RG/09/12/28096 (http://www.bhf.org.uk; to AR and WJA). KF and MK were supported by Marie Curie funding from the NETSIM FP7 program funded by the European Commission. The laboratory receives funding from the NHS Blood and Transplant for facilities. The Cambridge BioResource (http://www.cambridgebioresource.org.uk), the Cell Phenotyping Hub, and the Cambridge Translational GenOmics laboratory (http://www.catgo.org.uk) are supported by an NIHR grant to the Cambridge NIHR Biomedical Research Centre (BRC). The BRIDGE-Bleeding and Platelet Disorders Consortium is supported by the NIHR BioResource - Rare Diseases (http://bioresource.nihr.ac.uk/; to ET, NF, and Whole Exome Sequencing effort). Research in the Soranzo laboratory (LV, NS and SW) is further supported by the Wellcome Trust (Grant Codes WT098051 and WT091310) and the EU FP7 EPIGENESYS initiative (Grant Code 257082). Research in the Cvejic laboratory (AC and CL) is funded by the Cancer Research UK under grant number C45041/A14953. SJS is funded by NIHR. MEF is supported by a British Heart Foundation Clinical Research Training Fellowship, number FS/12/27/29405. EBM is supported by a Wellcome Trust grant, number 084183/Z/07/Z. Research in the Laffan laboratory is supported by Imperial College BRC. FAC, CL and SW are supported by MRC Clinical Training Fellowships and TB by a British Society of Haematology/NHS Blood and Transplant grant. RJR is a Principal Research Fellow of the Wellcome Trust, grant No. 082961/Z/07/Z. Research in the Flicek laboratory is also supported by the Wellcome Trust (grant number 095908) and EMBL. Research in the Bertone laboratory is supported by EMBL. KF and CvG are supported by FWO-Vlaanderen through grant G.0B17.13N.

PF is a compensated member of the Omicia, Inc Scientific Advisory Board.

This study made use of data generated by the UK10K Consortium, derived from samples from the Cohorts arm of the project. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org

The authors wish to acknowledge the following facilities and people:
Cambridge BRC Cell Phenotyping Hub, Cambridge BRC Translational Genomics Laboratory, Cambridge BRC Blood and Stem Cell Biobank, NHS Blood and Transplant donation clinic, Sanger Institute-EBI Single-Cell Genomics Centre, J. Garwood, D. Fletcher, L. Guglielmi and team of the Ouwehand lab, M. Bussemakers from the University of Nijmegen. The critical comments from Prof. J.-P. Allain and Dr. D. Rico were greatly appreciated.

The data are available from the Blueprint Consortium website http://www.blueprint-epigenome.eu/ and the European Genome-phenome Archive with accession number EGAD00001000745. Visualization tools are also accessible from http://blueprint.haem.cam.ac.uk/.

Fig. 1. Transcriptional atlas of hematopoietic progenitors and precursors. (A) Schematic representation of the current model of hematopoietic cell ontogeny and samples used in this study. Established ontological relationships

are represented as solid lines, emerging ontological relationship are represented as dotted lines. A simplified representation of mature cells is shaded. Antigens used for selecting each population are also indicated. The bone marrow residing components are the hematopoietic stem cell (HSC, light blue), multipotent progenitor (MPP, dark blue), lymphoid-primed multipotent progenitor (LMPP), common lymphoid progenitor (CLP, light green), common myeloid progenitor (CMP, dark green), granulocyte monocyte progenitor (GMP, light red), megakaryocyte erythrocyte progenitor (MEP, red), erythroblast (EB, light orange), megakaryocyte (MK, orange). The blood residing components are platelets (P), erythrocyte (E), neutrophil (N), eosinophil (Eo), monocyte (M) and lymphocyte (L). (B) Data analysis strategy. Reads were mapped to the transcriptome to quantify expression at the gene and transcript levels as well as the transcript proportion (defined as the fraction of gene expression level from a given transcript). Complementary to that, reads were mapped to the genome to identify novel splice junctions and sites where alternative splicing occurs. (C) Schematic highlighting the difference between assessing differential expression by looking at transcript expression or at transcript proportion.

Fig. 2. Transcriptional changes at lineage commitment events. (A) River plot representing gene expression levels across cell types for key transcription factors (TFs) required for lineage commitment. Line width represents expression level in log2(FPKM+1) normalized to the highest expression per gene across cell types. The relative changes in gene expression recapitulate the current understanding of the role of these TFs in hematopoietic differentiation. (B) Summary of the number of transcriptional classes - genes, transcripts and transcript proportions - changing at each lineage commitment point. Bayesian polytomous analysis was used to classify these 3 quantities into 5 possible models, from top to bottom: NULL model (no change); three single models (only one cell type different); and a FULL model (all three estimates differ). The number of events up or down regulated were tallied only when the change occurred in at least two samples at each branching event with an expression FPKM >1. (C) Cell-specific enrichment of protein-coding and non-protein coding biotypes in up and down regulated transcripts for the polytomous models at each branching event. (D) Heatmap of expression of lineage specific transcripts. Polytomous analysis was used to identify genes that were expressed significantly higher in a given cell type relative to all others. Top 20 highest scoring transcripts based on the posterior probability of the model are displayed. The colors along the left axis reflect whether the gene is protein coding (green) or otherwise (lilac).

Fig. 3. Cell-type specific splicing and RNA-binding motif enrichment in hematopoiesis. (A) Distribution of splice junction definition, absolute count and cell-type-specific fractions within unannotated splice junctions. Blue: annotated exons and junctions; Red: unannotated exons and junctions. (B) Cell-type specificity of known, unannotated and novel splice junctions measured with Shannon's entropy (*37*). Lower entropy indicates that splice junctions are observed in a smaller number of cell types. (C) Region-specific patterns of RNA-binding protein motifs around spliced-in and spliced-out DSU cassette exons. The enrichment or depletion of motifs in three regions: the 300bp intronic region adjacent to the upstream of the 5' splice site (blue); the exonic region of the cassette exon (orange); and the 300bp intronic region adjacent to the downstream of the 3' splice site (green). The heatmaps present significant enrichment (yellow) or depletion (red) in -log10 P value, FDR< 0.05.

Fig. 4. A novel isoform of the transcription factor NFIB regulates megakaryopoiesis. (A) A novel TSS and novel exon of *NFIB* was detected using RNA-seq (blue) and validated using 5' race PCR (red) and PacBio sequencing (green). Ensembl annotated transcripts in black. (B) Cartoon representation of the short and long isoforms of *NFIB* (*NFIB-S* and *NFIB-L*) highlighting the functional domains. (C) Western blot (WB) for NFIB, NFIC and Tubulin in megakaryocytes (MK), erythroblasts (EB) and monocytes (M) confirms that the short form of NFIB (NFIB-S) is predominantly expressed in MKs (* is either the protein product of one of the shorter transcripts of NFIB observed in the 5' race, or is unspecific). (D) Coimmunoprecipitation of overexpressed combinations of NFIC-HA together with TAP (Flag plus CBP) tagged NFIC, NFIB-L and NFIB-S. The upper panel was probed with anti-NFIC antibodies, showing both NFIC TAP tagged (upper band) and NFIC-HA tagged (lower band); note the absence of NFIC-HA in lane 4 showing lack of interaction between NFIC and NFIB-S. The lower panel was probed with anti-Flag antibody (part of the TAP tag), showing the immunoprecipitated NFIC (lane 2), NFIB-L (lane 3) and NFIB-S (lane 4) (see also Figs. S30 and S31), (E). Flow cytometry dot plots of CD41a and CD61 staining of megakaryocyte cultures at day 10 after infection with shRNA of control, NFIB and NFIC. The proportions of double positive, upper right (megakaryocytic), versus double negative, lower left (undifferentiated), cells decreased relative to control shRNA by silencing either NFIB or NFIC. (F) Overexpression of NFIC or NFIB-S lead to a higher proportion of megakaryocytic cells relative to NFIB-L or control. CD41a and CD42b double positive MKs in cultures at day 10 after infection. The y-axis is the probit proportion of double positive MKs after adjusting for batch effects.