Science AAAS

6 June 2014

SUPPLEMENTARY ONLINE MATERIALS

Companion to the paper:

”Transcriptional diversity during lineage commitment of human blood progenitors”

Index

Cord blood collection

Cord blood was collected with ethical approval (REC 12/EE/0040) after informed consent, according to the declaration of Helsinki, at the Rosie Maternity Hospital, Cambridge University Hospitals. Blood from each cord was processed within 18 hours of collection.

Isolation of cord blood progenitor cells

CD34^{+} progenitor cells were isolated from cord blood using the EasySep® progenitor cell enrichment kit with platelet depletion, according to manufacturer instructions (STEMCELL, Stemcell Technologies, Vancouver, British Columbia, Canada). Isolated CD34^{+} cells, from up to three cord blood units, were pooled and stained using a panel of monoclonal antibodies designed to identify six progenitor cell types.

Antibody cocktail:
PECy5^{-} lineage specific markers: CD2, Becton Dickinson (BD; NJ, USA) 555328; CD3, BD 561006; CD11b, BD 561686; CD14 eBIOSCIENCES (San Diego, CA, USA) 15-0149-42; CD19, BD 555414; CD56, BD 555517; GPA International Blood Group Reference Laboratory (IBGRL) 1/100
CD38 FITC (BD 560982)
CD34 APC (BD 560940)
CD90 PECy7 (BD 561558)
CD45RA Pacific Blue (Invitrogen, Life Technologies, Carlsbad, CA, USA; MHCD45RA28)
CD123 PerCP (BD 560904)
CD10 PE (BD 561002)

Stained cells were flow sorted using a BD fluorescence-activated cell sorting (FACS) ARIA III into six progenitor cell populations defined as:
HSC: Lin^{-} CD34^{+} CD38^{-} CD90^{+} CD45RA^{-}
MPP: Lin^{-} CD34^{+} CD38^{-} CD90^{-} CD45RA^{-}
CLP: Lin^{-} CD34^{+} CD38^{+} CD10^{+} CD45RA^{+}
CMP: Lin^{-} CD34^{+} CD38^{+} CD123 low CD45RA^{-}
MEP: Lin^{-} CD34^{+} CD38^{+} CD123^{-} CD45RA^{-}
GMP: Lin^{-} CD34^{+} CD38^{+} CD123^{+} CD45RA^{+}

Examples of the gating strategy used to sort the highly purified cell subsets are presented in Fig. S1. Sorted cells (20,000 to 200,000; >95% purity) were pelleted by centrifugation and lysed in TRIZOL reagent (Life Technologies, Carlsbad, CA, USA).

Erythroblast and megakaryocyte cell culture and purification

EBs and MKs were cultured from CD34$^+$ cells isolated from cord blood mononuclear cells using the human CD34 isolation kit (Miltenyi Biotec, Bergisch Gladbach, Germany) with an autoMACS separator (Miltenyi).

For EBs, cultures of 10^5 cells/ml were supplemented with 5 U/ml erythropoietin (EPO, R&D Systems, Inc. 614 McKinley Place NE, Minneapolis, MN 55413, USA), 100 ng/ml SCF (R&D) and 10 ng/ml IL3 (R&D) in CellGro media. Every 3 days, cultures were diluted 1:3 and media and cytokines rejuvenated. At day 14 mature EBs were immunopurified using an anti-CD36 PE conjugated antibody (BD Pharmingen, BD Biosciences, Applied Biosystems, Life Technologies, Carlsbad, CA, USA, 555455) and PE positive selection kit. The EBs purity was verified by FACS and shown to be >95% positive for CD36 (BD, 555454), CD71 (BD Pharmingen) and CD235a (IBGRL).

The method used to differentiate CD34$^+$ hematopoietic stem and progenitor cells into MKs, has been reported previously (51). MK cultures of 10^5 cells/ml were supplemented with 100 ng/ml thrombopoietin (rhTPO CellGenix, Freiburg, Germany) and 10 ng/ml IL1β (R&D) in CellGro media (CellGenix). Media and cytokines were rejuvenated at days 3 and 6. At day 10, mature MKs were immunopurified using an anti-CD42b PE conjugated antibody (Pab5, NHS Blood and Transplant, IBGRL) and a PE positive selection kit (STEMCELL). The MKs purity was verified by FACS and shown to be >95% positive for the surface markers CD41a (BD, 559777) and CD42b (IBGRL).

RNA-seq library preparation and sequencing

RNA was extracted from TRIZOL reagent using the manufacturer's instructions. Non-strand specific poly-(A)$^+$ RNA libraries were prepared using the SMARTer Ultra Low RNA and the Advantage 2 PCR kits (Clontech Laboratories, Inc., CA, USA), following the manufacturer's instructions and using 100 pg of total RNA as input. Libraries were indexed using NEXTflex adapters (Bioo-Scientific Corporation, Austin, TX, USA) and 100 base-pair (bp) paired-end sequencing was performed on Illumina HiSeq 2000 instruments using TruSeq reagents (Illumina, San Diego, CA, USA), according to manufacturer's instructions.

RNA-seq library sample exclusion

In total, four samples were removed from further analysis, three due to unacceptably high duplication rates and one because of spatial biases. We determined spatial biases present in the RNA-seq libraries since this precludes the comparison of transcripts and alternative splicing between multiple samples. To quantify the spatial bias per sample we summarized the median transcript coverage across all transcripts between 2,000 bp and 4,000 bp long and with a FPKM (fragments per kilobase per million reads) between 1 and 50. Coverage information was derived from the Bowtie (56) alignments to the transcriptome used in the MMSEQ tool (26) (cf. Transcriptome alignment and quantification of expression). We then compared the distributions for these transcripts using the Euclidean distance. The result of this clustering highlighted one of the samples as having a strong 5' underrepresentation.

RNA-seq adapter trimming

Paired-end reads were trimmed for both PCR and sequencing adapters using Trim Galore! v0.2.8 (55) with parameters `-q 15 -s 3`. For the first cycle of trimming, we also used parameter `-e 0.05` and retained reads of length equal to or larger than 36 bp. For the second cycle of trimming, parameter `-e` was set to its default value (`0.1`) in order for poly-(A)$^+$ tails to be successfully trimmed. We kept reads with a length equal to or larger than 20 bp. Unpaired reads were discarded during both trimming cycles.

Transcriptome alignment and quantification of expression

Trimmed reads were aligned to the Ensembl v70 human transcriptome using Bowtie v0.12.8 (56). The number of aligned reads is summarized in Table S1. We then used the MMSEQ v1.0.5 (26) tool to quantify gene and transcript isoform expression.

Pairwise differential expression analysis

Differential expression at the gene, transcript, and the isoform usage ratio was determined using MMDIFF v1.0.5 (30). The Bayesian linear mixed effects model, used by MMDIFF, takes into account the error estimates of the expression values quantified using MMSEQ. Therefore, MMDIFF down-weights poor estimates which may arise from transcripts that are difficult to deconvolute or from estimates from less deep libraries. For all differential expression analysis performed at the HSC to MPP transition, we performed a two-model comparison at the branching points. Under the first model, the mean expression level is the same in all cell types and under the second model the mean expression level is allowed to differ between cell types. The prior probability of the second model was set to 0.1. Features with a posterior probability of the second model greater than 0.5 were considered as differentially expressed. Results were filtered by requiring an FPKM>1 in at least two samples.

Polytomous model comparison

Polytomous model comparison was performed to determine the most likely direction of change of expression at branching points. Five models were compared. 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. Bayes factors (B) between the simplest model and each relaxed model were calculated and Bayes' theorem was used to compute the posterior probability that the true model γ is equal to m under the assumption that the five models are exhaustive:

$$P(\gamma = m|y) = \frac{B(0,m) \times P(\gamma = m)}{\sum\limits_{m'} B(0,m') \times P(\gamma = m')}$$

where y denotes the MMSEQ log expression estimate and model 0 is the simplest model. The prior for the simplest model was set conservatively to 0.8 and the remaining 0.2 prior belief was distributed equally across the four relaxed models. We also modeled posterior summaries of probit-transformed isoform usage proportions to detect transcript isoform switching events during differentiation. Specifically, at each iteration of the MMSEQ Markov chain Monte Carlo (MCMC) algorithm, the expression due to each transcript is divided by the total expression for the corresponding gene, probit-transformed and recorded. The empirical mean of the resulting trace is then treated as an outcome in the MMDIFF linear mixed model, where the empirical standard deviation is treated as the variance of a corresponding random effect.

Validation of transcript isoform quantification by qPCR

To validate the quantification of transcript levels determined by analysis of the RNA-seq data using MMDIFF, we performed quantitative PCR (qPCR) using 40 transcript specific assays and five positive control assays in multiple progenitor cell subsets. To simplify assay design, transcripts were selected to represent an exon-skipping event where there was a maximum of four transcript isoforms annotated for the gene. Transcripts were also selected to represent variable expression over the multiple cell types. qPCR primer and probe sets were designed to be specific for a transcript and primers and probes, FAM labeled and quenched using Tamra, were manufactured by Sigma-Aldrich (Haverhill, UK) (Table S2). A 20× working mix of primers and probes was generated (6 μM each primer and 4 μM probe).

RNA from 16 independent samples from 5 progenitor cell subsets was purified as described above (cf. RNA-seq library preparation and sequencing) and cDNA generated using SuperScript VILO cDNA Synthesis Kit (Invitrogen), according to the manufacturer's instructions. cDNA was pre-amplified and then amplified using TaqMan PreAmp Master Mix and TaqMan Gene Expression Master Mix respectively, using the manufacturer's protocols (Applied Biosystems, Life Technologies). Quantification of transcripts was performed in each sample, in duplicate, using the BioMark system (Fluidigm Corporation, San Francisco, CA, USA). After requiring call quality scores >0.9 (Fluidigm Real-Time PCR Analysis software), 36 transcripts were analyzed. Cq values were determined using the Fluidigm Real-Time PCR Analysis software and values exported for statistical analysis.

Normalization of the Cq values was performed using the $\Delta\Delta$Cq method. First, by subtracting the expression of *B2M* Cq values (similar results were achieved for *ACTIN B*). Next, by subtracting the average ΔCq for that probe in MKs. Linear regression was performed between the $\Delta\Delta$Cq values and the RNA-seq transcript average for the given cell type, excluding the values from MK (Fig. S2).

Biotype and gene ontology enrichment analysis

Gene ontology enrichment analysis was performed using the `goseq` R package (`http://www.bioconductor.org/packages/release/bioc/html/goseq.html`) that accounts for the dependency between power to call differential expression and gene length and thus corrects for selection biases (57). All P values were corrected for multiple testing using the Benjamini-Hochberg method. Transcript and gene biotypes were retrieved from Ensembl v70 (Fig. 2C).

Cell-type specific genes and transcripts

The selection of cell-type specific genes and transcripts was based on a 9-model polytomous classification and a set of stringent thresholds on expression fold changes and posterior proba-

bilities. The 9 models include the null model, under which expression does not differ between cell types, and 8 alternative models, each representing expression diferences in one cell type against the pool of remaining cell types. The null model was assigned a prior probability of 0.5 and all others of 0.5/8. To select cell-type specific markers we then required that the posterior probability of the alternative model was greater than 0.5, that the gene/transcript was expressed at and FPKM>0 in all three replicates of a cell type and that the $\log_2$ fold expression was higher than 1 in a cell type versus the pool average.
To compare the gene expression of cell-type specific genes between our RNA-seq dataset and publicly available microarrays, probe annotations for Illumina (16) and Affymetrix (Santa Clara, CA, USA, 14) platforms were retrieved from Ensembl v70 (Fig. S5 A-D). Cell-specific genes derived from the RNAseq data were tested for enrichment in the array data using a two-sided Wilcoxon rank sum test. The normalized (per progenitor cell) test statistics were used to display the degree of enrichment. This comparison demonstrates that our RNA-seq cell-specific gene signatures agree with microarray findings. The results for CMP-specific gene sets were not significantly similar to any cell population, likely due to differences in gating strategy for the expression of CD123/IL3Ra (low vs. absent, Fig. S1).

Genome alignment to identify splicing events

Trimmed reads were aligned to the February 2009 (GRCh37) version of the reference human genome using three different aligners:
GSNAP v2013-10-15 (59), STAR v20201 (60) and GEM tools v1.6.2 (61).
GSNAP: Read trimming was disabled, a maximum of 7 mismatches and 10 multi-hits were allowed, and novel splice sites were limited to a maximum of 300 kb apart.
STAR: Default settings were used as they were optimized for 75-100 bp paired-end reads in human.
GEM: We used a quality offset for the FASTQ score of 33, mismatches were allowed up to a maximum fraction of 0.06, and a maximum of 10 hits for the multi-mapping reads were allowed.

For both GSNAP and STAR alignments, annotated splice junctions retrieved from Ensembl v70 were provided.

Identification of novel and alternative splicing events

Splice junctions were extracted from BAM files generated by GSNAP (59), STAR (60) and GEM (61). We required splice junctions to be supported by at least 10 reads and for each aligned read to span at least 10 bp at each side of the splice junction. To reduce potential biases introduced by the aligners, splice junctions reported by all three aligners in a minimum of two samples were used for downstream analysis (Fig. S6).

Unnanotated splice junctions were identified through the comparison of our set to those annotated in Ensembl v70. Comparison to the most recent Ensembl release (v75) accounted for 420 additional splice junctions. This analysis identified a number of splicing events that were shifted by 1 bp from annotated splice junctions. Motif analysis confirmed that 99.6% of these did not follow the nucleotide patterns required for U2 splicesome binding (GT-AG or GC-AG), therefore we discarded these from further analysis.

Additional validation of splice junctions included two distinct datasets: (i) EST and mRNA data downloaded from the UCSC Genome Browser (63) and (ii) the Illumina BodyMap 2.0 dataset (35), which is a set of RNA-seq data from 16 human individuals and a series of mixed tissues (GEO accession: GSE30611).

A total of 8,704,794 ESTs and 1,976,868 mRNAs were downloaded from the UCSC Genome Browser (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/`) on 2nd April 2013 and aligned to the reference human genome (GRCh37) using GMAP (64). From this alignment, we retrieved 477,711 unique splice junctions. The Illumina BodyMap 2.0 dataset was re-aligned to the reference human genome (GRCh37) using GSNAP, identifying 1,234,965 unique splice junctions. To identify alternative splicing events, we examined splice junctions that share either the acceptor or the donor site. These splicing events were classified as either exon-skipping, alternative 3' splice site, or alternative 5' splice site by comparison to Ensembl v70 annotations.

Characterization of novel splice junctions

Splice site probability scores in Fig. S7 were extracted from the GSNAP output using the default output format. To perform motif analysis on the splice sites, a 5 bp sequence around each splice site (2 bp to the intron and 3 bp to the exon) was considered (Fig. S11).

The evolutionary conservation around all splice junctions was examined based on multiple alignments from vertebrates and primates, using phastCons conservation scores (65) (`http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/`).

We compared phastCons conservation scores in a 100 bp window, 50 bp on each side of the splice junctions (Fig. S8). To assess the coding potential of the sequence within a 100 bp window, we summed all stop codons in all six possible reading frames (three on each strand) in both exonic and intronic regions. A lower number of stop codons is indicative of a higher potential to transcribe a protein-coding transcript. Unannotated and known coding regions had similar distributions of stop codons, whereas intronic regions had higher numbers of stop codons, suggesting that the unannotated exons we identified have similar coding potential as known exons (Fig. S9).

Shannon's entropy was calculated for annotated, unannotated and novel splice junctions (37) based on the read coverage of the splice junctions. A lower entropy indicates that splice junctions are relatively highly expressed in a small number of cell types, as also observed in a heatmap of read counts of novel splice junctions (Fig. S12).

Validation of novel splice junctions by PCR

To validate progenitor-specific novel splice junctions we designed PCR primers for 23 exon-exon events (Table S9). Primers were also designed to amplify regions of the Actin B and Albumin genes as positive and negative controls, respectively, for the input material. Water controls were run alongside all PCRs. Illumina RNA sequencing libraries were pooled for each cell type (HSC, CLP, MEP, EB and MK) and subjected to a further round of 18 cycles of PCR amplification using Illumina primers. Each PCR reaction contained 25 ng of Illumina library DNA template and 10 pmol of each gene specific primer in a PCR master mix (Bioline, London, UK) with a final volume of 50 µl. PCR products were visualized on a 2% TBE/agarose gel and scored by eye. PCR products with a single band and sufficient sample concentration (15 out of 23) were sequenced using Sanger sequencing. These were aligned to the human genome (GRCh37) using BLAT in the UCSC Genome Browser (`http://genome.ucsc.edu/cgi-bin/hgBlat?command=start`). The 23 primer pairs flanked novel splice junctions identified in HSCs, CLPs, MEPs, EBs and MKs, and included nine novel splice junctions, nine novel acceptor sites, three novel donor sites and two DSU exon-skipping events (cf. Identification of differential splice usage) (Fig. S13-S15, Table S10).

For each PCR validation of a novel splice junction we have displayed the RNA-seq reads and gel pictures of the PCR products (Fig. S9). RNA-seq reads and splice junctions were visualized using IGV (`http://www.broadinstitute.org/software/igv/home`) and sashimi plots (`http://genes.mit.edu/burgelab/miso/docs/`

sashimi.html). Fig. S13L displays a novel acceptor site at an intronic region in *SYNGR1*. Primers for the splice junction amplified a specific product in EBs (blue arrow), which is concordant with the RNA-seq data. However, a non-specific band is observed in MEPs.

We validated two splice junctions where our RNA-seq DSU analysis supported a cell specific exon skipping event (cf. Identification of differential splice usage) (Fig. S14). At an intronic region of *NUP155* on chromosome 5 we identified a CLP-specific DSU event in our RNA-seq data with an FDR = 2.15×10^{-5} (Likelihood ratio test) (Fig. S14A). We observed an additional PCR amplicon ($\sim$200 bp) in the CLP sample, confirming this finding (Fig. S14A). We also identified an EB-specific DSU exon-skipping event (FDR = 9.98×10^{-4}, Likelihood ratio test) in CD71 (transferrin receptor) that we confirmed using PCR (Fig. S14 B).

Using PacBio (Pacific Biosciences, Menlo Park, CA, USA) data (cf. Validation of transcript isoform detection using PacBio) we could identify novel splice junctions and novel exons within full-length transcripts within MKs. As this data is strand-specific, in the following examples, we have identified the transcripts containing these novel exons and splice junctions.

Firstly a CLP-specific novel splice junction was identified spanning two novel exons. The 5' novel exon was within the MTHFD2L gene and the 3' novel exon downstream of *EPGN*. Using PCR we confirmed this splice junction, and thus these exons, in CLPs (Fig. S15A). Concordant with the RNA-seq data, no PCR products were detected in EBs and MKs. In our MK PacBio data we identified transcripts containing the novel exon downstream of EPGN (Fig. S15A, lower panel). The strand specific PacBio data suggests this novel exon is part of an unannotated transcript on the reverse strand that is differentially spliced in MKs and CLPs.

A novel splice junction spanning two novel exons was identified within an intergenic region of chromosome 12. Using PCR we validated the splice junction in HSCs, CLPs, MEPs, EBs and MKs (Fig. S15B). The alternative acceptor site (within the left exon) was identified in the RNA-seq, PCR data and PacBio data (Fig. S15B, lower panel). The transcript with the 5' alternative splicing event, corresponding to the short PCR product, was also identified within the PacBio data for MKs.

We identified two novel splicing events spanning two novel exons within an intronic region of GNG12 (Fig. S15C). Using PCR we validated the first splicing event in CLPs, EBs and MKs (205 bp product). In EBs and MKs we validated the second alternative acceptor site identified in the RNA-seq data ($\sim$120 bp product). Within the PacBio reads from the MK, both of these novel exons were observed, however, within transcripts on the forward strand. This result suggests that these novel exons are not within transcripts of *GNG12*, which is transcribed from the reverse strand (Fig. S15C, lower panel).

In addition to validating the novel, cell-type specific DSU events, an additional 7 PCR assays were designed for known DSU events (Fig. S23). We looked to validate the quantitation of the PSI estimation from the RNA-seq by comparing with the PSI calculated using the PCRs. Using a densitometer we quantified each band observed in these PCRs and where DSU was observed calculated the PSI. The PSI of 26 DSU events, in 11 PCR assays, correlated with the PSI obtained by RNA-seq (R^2 = 0.78, Fig. S24).

Validation of transcript isoform detection using PacBio

Megakaryocyte total RNA (10 ng) was reverse transcribed and amplified using the SMART-Seq2 protocol (66). cDNA was pre-amplified and purified using SPRI (Beckman Coulter Genomics, Danvers, Massachusetts, USA) beads at a 1:1 volumetric ratio. Eight individual reactions from the same starting material were performed and the resultant cDNA pooled. SMRTbell libraries were made and bound to P4 polymerase as per existing PacBio protocols for a 2 kb library (DNA Template Prep Kit 2.0 (from 250 bp to 3 kb) and DNA/Polymerase Binding Kit P4). Bound complexes were loaded on V3 SMRTcells using MagBeads, and sequenced using C2 chemistry (DNA Sequencing Kit 2.0) and 180 min movies.

We sequenced five SMRTcells (Pacific Biosciences, Menlo Park, CA, USA) in total. Primary filtering of the sequence data showed ∼1.8 Gb of total data, consisting of 435,366 reads with a mean polymerase read length of 4.1 kb. Secondary analysis, using SMRT-pipe version v2.1.0, generated the following sub-read information; 1,004,331 sub-reads, mean sub-read length of 1.7 kb, N50 = 2.1 kb. Read-of-insert (ROI) filtering, creating the best consensus-read from the multiple passes (sub-reads) of each SMRTbell molecule, provided 399,150 ROI with a mean length of 1.962 kb (median, 1.821 kb) which were used for subsequent transcript analysis. After removal of chimeric reads, 387,179 non-chimeric ROIs were analyzed for the identification of full-length transcripts. In total, 67,110 ROIs were identified as full-length reads, containing the 5’ and 3’ adapters and a poly-(A/T) tail. These were clustered based on their sequence similarity to 35,663 consensus sequence clusters using ICE v1.1.6. A total of 35,476 consensus sequence clusters were uniquely mapped to the reference human genome (GRCh37) using GMAP v2014-04-21 with parameters `--format=2 -t 24 -n 1 --nofails`.

All scripts used for the processing and analysis of the PacBio dataset can be found at `https://github.com/PacificBiosciences/cDNA_primer`.

Identification of differential splice usage

A beta binomial model was used to identify differential splicing usage (DSU) events, accounting for the overdispersion in the read count data. Only alternative splicing sets with two splicing events were used. Multinomial Dirichlet models were considered for splicing sets with more than two options but often displayed poor convergence properties with our data. To fit the beta binomial model we used the `vglm` function from the `VGAM` R package (`http://CRAN.R-project.org/package=VGAM`). The model estimates the usage proportion (the probability parameter of the binomial distribution) of a given splicing alternative (the shortest one). P values were determined using likelihood ratio tests comparing the model with a factor accounting for difference in one cell type versus a pool of all others against a single common usage proportion. Samples where no reads were observed for either splicing event were ignored in the fit. P values were corrected for multiple testing using FDR. For each splice junction a single *percent-spliced-in* (PSI) was calculated using regression parameters. Using the beta binomial model to account for overdispersion led to a more uniform distribution of the P values under the null distribution, and thus better calibrated P values, compared to alternative models.

In GFI1B, an exon-skipping event was identified with two alternative splice junctions (ASJ1 and ASJ2 from left to right, note the common 5' coordinate). The percentage of ASJ2 was increased in CMPs while being absent in MEPs and GMPs. The proportions of ASJ1 and ASJ2 significantly changed among cell types (P=0.0021 and 0.0012 in GSNAP and STAR, respectively) (Fig. S19).

Gene ontology enrichment analysis was performed on all genes with DSU (Table S12). To examine the consequences of exon-skipping events on protein domains, we used InterProscan 5 (`http://www.ebi.ac.uk/Tools/pfa/iprscan5/`) to search for domains predicted by Pfam (`http://pfam.xfam.org/`) on cassette exons with DSU.

Enrichment analysis of RNA-binding motifs around up- and down-regulated cassette exons

To assess the cell-type specificity of the binding of splice factors around cassette exons with DSU, we performed motif enrichment analysis by adapting a method previously published by Castle et al. (67). We used a set of 80 RNA-binding proteins, represented by a total of 102 position weight matrices (PWMs) (41). These PWMs were filtered for motifs with high motif similarity, as identified using Tomtom (`http://meme.nbcr.net/meme/doc/`

tomtom.html) (Table S13).

For all cassette exons, we extracted the nucleotide sequence in three regions; the 300 bp intronic region adjacent to the upstream 5' splice site; the exonic region of the cassette exons; and the 300 bp intronic region adjacent to the downstream 3' splice site.

For each cell type, the test set included up- and down-regulated cassette exons identified with an increased or decreased percentage of splice in (PSI) $> 5\%$, respectively, compared to all other cell types and a FDR < 0.1 based on the DSU analysis. The background set included all cassette exons.

Cell-type specific enrichment analysis of each PWM was calculated using the cumulative hypergeometric distribution. Importantly, we used the number of motif occurrences in the DSU set rather than the presence or absence of at least one motif. Two-tailed P values were calculated as $2 \times \min(P, 1-P)$. The P values that were obtained represent the probability of observing the number of occurrences in the test set versus the number in the background. P values were corrected for multiple testing using FDR and assigned positive for enrichment or negative for depletion. For RNA-binding proteins represented by more than one motif, we retained the motif with the lowest P value.

Chromatin immunoprecipitation and sequencing

Chromatin immunoprecipitation and sequencing (ChIP-seq) for MEIS1 was performed on CD34$^+$ *in vitro* derived MKs as previously described (51, 68). ChIP-seq libraries were prepared according to the Illumina ChIP-seq preparation kit.

Sequencing was performed on a HiSeq 2000 machine, producing two lanes of 54 bp single-end reads. Read quality assessment was done using FASTQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to the reference human genome (GRCh37) using Bowtie (56). Post-alignment quality control was based on the IP enrichment, as previously described (69). Identification of enriched regions was performed using MACS v1.4 (http://liulab.dfci.harvard.edu/MACS/) without a control sample.

Cloning, shRNA, lentivirus production and transduction

TRC shRNA lentivirus targeting *NFIB* and *NFIC* and a non-silencing control were purchased from Thermo Scientific Open Biosystem (RHS4080; RHS3979-201746256-9 and RHS3979-201746286-90). cDNA was either purchased from Thermo Scientific Open Biosystem (GE

Healthcare, Little Chalfont, UK) (*GFI1B*, *NFIB* full length and *NFIC*) or cloned from MK cDNA libraries (*NFIB* short). *GFI1B*, *NFIB* full length and *NFIC* cDNA were cloned, sequenced and then subcloned in pWPI (D. Trono lab) TAP tagged (calmodulin binding protein + flag).

Packaging was performed in 293T cells and viral stocks were titrated and quantified using qPCR and for pWPI also using GFP FACS. A Multiplicity of Infection of 50 was used in each infection.

CD34$^+$ cells were purified as above (Miltenyi) from NHSBT mononuclear blood cell concentrate. This cell fraction remains in the filter cone during a platelet apheresis procedure. All NHSBT donors consent for the use of surplus from their platelet or blood donation to be used in an anonymized manner for biomedical research. CD34$^+$ cells were infected with lentiviral particles in the presence of polybrene (8 μg/ml), overnight in CellGro supplemented with 100 ng/ml TPO and 10 ng/ml IL1β.The following day media was replaced and the cells allowed to differentiate into MKs (cf. Erythroblast and megakaryocyte cell culture and purification). At day 10, MKs were counted and assessed by morphology and flow cytometry for maturation using CD41a (BD, 559777) and CD42b (IBGRL).

Transfections, immunoprecipitations and Western blots

To detect protein-protein interactions, NFI proteins were expressed by co-transfection in 293T cells using PEI (1 mg/ml, 25,000 MW, Polyscience Inc., 400 Valley Road, Warrington, PA, USA). Cells were collected after 48 hours and lysed in Flag M2 lysis buffer (Sigma-Aldrich) according to manufacturer's recommendations. M2 resin (Sigma-Aldrich) was added overnight and after extensive washes, bound proteins were eluted in Laemmli SDS buffer (Sigma-Aldrich), boiled and separated by SDS Page (Nupage, Invitrogen). After transfer to polyvinylidene difluoride membrane, Western blots were performed using anti-Flag antibodies (M2 Sigma-Aldrich) and anti-NFIC antibodies (Bethyl laboratories A303_123A) with anti-mouse or anti-rabbit peroxidase conjugated secondary antibodies for final detection (Pierce, Thermo Scientific).

Knock down of NFIB and NFIC using shRNAs were successful and did not alter the endogenous level of the other protein. The overexpression of NFIC and both the long and short isoforms of NFIB was observed in K562 cells and did not alter the endogenous levels of NFIB and NFIC (Fig. S30 and S31).

Distinct functions of GFI1B isoforms in normal and pathological megakaryopoiesis

A gene highlighted in both the polytomous transcript and DSU analyses is *GFI1B*. GFI1B is a transcriptional repressor and oncogene in hematopoiesis with an essential role in the regulation of megakaryopoiesis and erythropoiesis (70). The polytomous analysis revealed differential transcript isoform usage at the MPPs towards the CMP branching point (transcript ENST00000372123) and the MEPs towards the MKs and EBs (transcripts ENST00000372123 and ENST00000534944) branching point (Fig. S32).

The six annotated protein-coding transcripts for GFI1B can be grouped into two sets on the basis of the length of the encoded protein isoform: a long (L) and short (S) isoform (GFI1B-L=330 aa CCDS293483; GFI1B-S=284 aa, CCDS03947; Fig. S33). The *GFI1B-S* transcript lacks exon 4, resulting in the exclusion of the first two of six zinc (Zn)-finger domains present in GFI1B-L (Fig. S33). Transcripts encoding *GFI1B-L* are expressed within all progenitor and precursor cells, with the exception of CLPs. Transcripts coding for *GFI1B-S* are only present in CMPs, EBs and MKs (Fig. S32) and their expression is elevated in myeloid leukemias compared to healthy controls (71). In human CD34 cells, overexpression of *GFI1B-L* promotes cell number expansion, but impairs erythropoietic terminal maturation (72). In contrast, in a cellular model of erythroid commitment, both *GFI1B* isoforms are required to promote differentiation (73). Coding dominant nonsense mutations in the fifth and sixth Zn-finger domains of GFI1B have been observed in subjects with atypical megakaryopoiesis (Fig. S33). Patients carrying these mutations display reduced numbers of platelets and a paucity of platelet alpha-granules (74,75). To better understand the role of *GFI1B* in megakaryopoiesis we reviewed the mutational spectrum at the *GFI1B* locus in 529 exome-sequenced cases with atypical platelet phenotypes enrolled by the BRIDGE consortium (cf. BRIDGE consortium). We identified seven patients with non-synonymous single nucleotide polymorphism (SNPs) absent in 11,216 unaffected individuals from reference cohorts (Fig. S32, Table S15). Four of seven of these non-synonymous SNPs are located in Zn-fingers 1 and 2. Patient BPD-B200597, of Asian Indian ancestry, carried a homozygous variant resulting in a Cys168Phe mutation (Fig. S34) absent from 321 Indian Asian genomes, and displays abnormal function and reduced number of platelets. Alignment of human *GFI1B* and 34 other eutherian orthologs showed sequence conservation across the eutheria for the six Zn-finger domains (residues 163–327) (Fig. S35). The structure of the Zn-finger of the homologous protein EGR1/Zif268 (PDB entry 1g2d) (76) suggested that the Cys168Phe mutation prevents the stable folding of the first Zn-finger domain of GFI1B (Fig. S34). All this evidence suggests the functional importance of the long form of GFI1B in human megakaryopoiesis and the relevance of its first Zn-finger domain. However, the physiological relevance of the alternative Zn-fingers in human megakaryopoiesis remains unresolved. To clarify the role of the two isoforms in megakaryopoiesis we overexpressed the L- and S- isoforms in $CD34^+$ cells. There was a significant expansion in the number of MKs for the L-form, whereas, no difference was observed for the S-form (Fig. S36). Neither isoform had an effect on MK maturation compared to the control vector (Fig. S37). The functional differences between the two *GFI1B* isoforms were confirmed by overexpression studies in a zebrafish model of thrombopoiesis where the human *GFI1B-L* mimicked the overexpression of the zebrafish ortholog of similar length, whilst the S-isoform had no effect (Fig. S38). Taken together, these results suggest that the two isoforms are functionally distinct and only GFI1B-L retains its functionality.

BRIDGE Consortium

The BRIDGE (Biomedical Research Centres/Units Inherited Diseases Genetic Evaluation) consortium (`www.bridgestudy.org.uk`) has been established as an organizational umbrella for the NIHR BioResource – Rare Diseases (`http://bioresource.nihr.ac.uk/`) sequencing projects. The participants of the BRIDGE-BPD consortium are listed in Table S14. The aim of the BRIDGE projects is to identify novel variants that underlie inherited rare disorders of unknown genetic etiology. To date, for the bleeding and platelet disorders (BPD) project of BRIDGE, 697 individuals, mainly probands, have been enrolled by 14 enrolment centres [UK (6), Belgium (1), France (2), Germany (2), USA (2), Australia (1), see Table S14].

All UK patients in the BRIDGE-BPD study were recruited with informed consent (UK Ethics number 04/Q0108/44) and for overseas samples informed consent was obtained under national ethics regulations pertinent to that country. The inclusion criteria for enrolment are: (i) positive history of bleeding, (ii) abnormal platelets (abnormal count [$<100 \times 10^9$/L or $>400 \times 10^9$/L] and/or volume [<6 Fl or >12 Fl] and/or function [abnormal functional response to any of the typical platelet agonists], and/or morphology), (iii) combination of (i) and (ii), (iv) high likelihood of being of genetic etiology (e.g. early onset, informative pedigrees, absence of acquired cause). Exclusion criteria are: (i) use of prescription or over-the-counter drugs known to be associated with abnormal platelet phenotypes or bleeding, (ii) high likelihood of autoimmune thrombocytopenia or other autoimmune disorders associated with low platelet count (including HIV positivity), (iii) other medical conditions known to be associated with acquired abnormal platelet phenotypes, e.g. malignancies, particularly those compromising hematopoiesis; bone marrow aplasia; thrombotic thrombocytopenic purpura or hemolytic-uremic syndrome; acute viral infection; splenomegaly; uremia or hepatic failure.

Exons have been captured using the NimbleGen SeqCap EZ Human Exome v3.0 chip (Roche, Basel, Switzerland) targeting 64 Mb. Captured exons and genomes have been sequenced using Illumina HiSeq instruments. Clinical phenotype information was captured and coded in the research database by assigning Human Phenotype Ontology (HPO) terms to each patient (77). Relevant laboratory information (e.g. full blood count, platelet function and coagulation tests, microscopic blood smear assessment, etc.) was also retrieved from clinical notes and coded. All information was deposited in the NIHR BioResource Rare Diseases study database.

DNA sequencing, alignment and variant calling in the BRIDGE project

Multiplexed DNA libraries were sequenced on Illumina HiSeq instruments. Adapters were trimmed using Trim Galore! v0.2.8 (55) and sequence read groups only passed quality control if the ”Per base sequence quality”, ”Per sequence quality scores” and ”Per base N content” FastQC v0.10.1 (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`) checks did not fail.

Alignment to the 1000 Genomes (78) build of GRCh37 was performed using BWA 0.6.2 (79). Realignment around indels and base call quality recalibration was performed using GATK 2.3_9 (80). Possible PCR duplicates were marked using Picard 1.89 (`http://picard.sourceforge.net`). We required that the percentage of target bases at or above 10× be greater than 70% and the estimated contamination rate be less than 5%, as estimated by VerifyBamID (81). Multisample genotype calling was performed using GATK and variant call qualities were recalibrated according to GATK best practices. Only genotype calls with a quality score greater than 20 were considered. Variants were annotated against allele frequencies and counts within the 1000 Genomes, UK10K (`www.uk10k.org/`) and ESP (`http://evs.gs.washington.edu/EVS`) datasets. None of the rare *GFI1B* variants reported here were present in any of the 11,216 samples in these reference datasets. Observed and reported variants (Fig. S33, Table S15) were confirmed by Sanger sequencing.

Overexpression of GFI1B transcripts in zebrafish

The maintenance, embryo collection and staging of the wild type (Tubingen Long Fin) and transgenic (Tg(cd41:GFP)) line was performed according to the Zebrafish Book (82). Zebrafish embryos were maintained at 28°C in egg water (60 mg/l Red Sea salts). Full-length zebrafish gfi1b cDNA was cloned into pCS2 expression vector using gene-specific primers:
GTTATAGAATTCATGCCACGGTCGTTTCTGGTG (forward)
AGGCGCCTCGAGTTATTTAAGGCTGTGCTGGCT (reverse)
Restriction enzyme sites (EcoRI/XhoI) used for cloning are underlined. Long human (h) *GFI1B*, short *hGFI1B*, zebrafish (zf)-*gfi1b* and memCherry (control) mRNA were synthesized with mMESSAGE mMACHINE kit (Ambion, Life Technologies), according to the manufacturer's protocol. 300pg of the long (L) and short (S) form of human *GFI1B* (*hGFI1B-L* and *hGFI1B-S*, respectively) and memCherry mRNA or 40 pg of (zf)-*gfi1b* mRNA were injected into the one cell stage Tg(cd41:EGFP) embryos. At three days post fertilization (dpf) embryos were anesthesized in 0.02% MS222 solution (Sigma-Aldrich) and images of the caudal hematopoietic tissue captured using a Leica TCS SP5 confocal microscope with Leica LAS AF software (Leica Microsystems Inc., Buffalo Grove, IL, USA), using a 40× immersion lens.

Evolutionary analysis of *GFI1B*

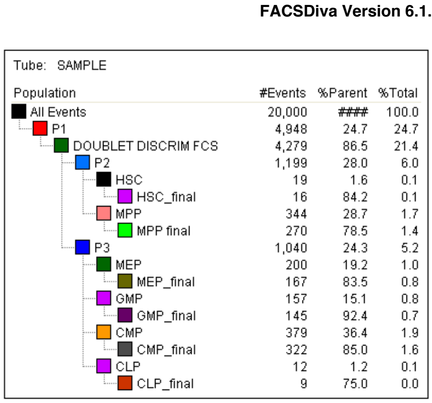
An evolutionary analysis was performed for the coding fraction of the *GFI1B* to better appreciate the consequences of the mutations observed in the seven BRIDGE-BPD patients (Fig. S35). Orthologs of human *GFI1B* were identified in 34 eutherians using the Ensembl Compara database, and coding DNA sequences for the canonical transcripts were downloaded from Ensembl release 74 using the `PyCogent` package (83). The corresponding gene tree was downloaded from Ensembl Compara and the 35 DNA sequences were aligned using PRANK (84) in codon mode. The resulting alignment was *humanized*, i.e. filtered to retain only positions present in the human sequence. Site-wise selective pressure was estimated using SLR with default parameters (85). The results obtained were plotted using the `PhyloSim` R package (86). Site-wise selective pressure was estimated using the relative non-synonymous:synonymous rate ratio (dN/dS) using `codeml` (87) (options `CodonFreq=2 [F3x4], model=0, NSsites=8 [beta&ω], ncatG=5`) and were found to be highly correlated with results from SLR (r=0.95). Finally site-wise evolutionary rates were estimated using baseml (87) using the general time reversible model with each codon position having independent evolutionary rate, substitution pattern parameters and Gamma-distributed rate heterogeneity (options `REV, Mgene=4, Malpha=1, ncatG=4`). For this sequence features were obtained from Uniprot (88).

The periodic pattern of increased rate at third codon positions, typical of protein-coding DNA, is clearly visible throughout *GFI1B*. In summary, the alignment reveals high sequence conservation across the eutheria. The SNAG domain in the N-terminal region (res. 1–20; exon 1) and the region harbouring the six Zinc-finger domains (res. 163–327; exons 3–6) are highly conserved. This is reflected both in the dN/dS and evolutionary rate estimates in these regions, indicating high evolutionary constraint. The region (approx. res. 21–97; exon 3) following the SNAG domain is less constrained, however still indicative of purifying selection. The region following this but before the zinc-finger domains (approx. res. 98–162; exon 3) is again very highly constrained. Frameshift mutations (marked with grey pointers) do not appear in the alignments and as such are not amenable to evolutionary analysis, however, 6 out of 7 remaining variants (orange pointers: G139S, C168F, H181Y, R184P, G198S, Q287STOP) fall on sites that are completely conserved on the amino-acid level across the eutheria, suggesting their importance for maintaining protein function.

Data access and visualization

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`.

Supplementary figures



FACSDiva Version 6.1.3

Tube: SAMPLE

Population	#Events	%Parent	%Total
⬛ All Events	20,000	####	100.0
🟥 P1	4,948	24.7	24.7
🟩 DOUBLET DISCRIM FCS	4,279	86.5	21.4
🟦 P2	1,199	28.0	6.0
⬛ HSC	19	1.6	0.1
🟪 HSC_final	16	84.2	0.1
🟥 MPP	344	28.7	1.7
🟩 MPP final	270	78.5	1.4
🟦 P3	1,040	24.3	5.2
🟩 MEP	200	19.2	1.0
🟫 MEP_final	167	83.5	0.8
🟪 GMP	157	15.1	0.8
🟪 GMP_final	145	92.4	0.7
🟧 CMP	379	36.4	1.9
⬛ CMP_final	322	85.0	1.6
🟪 CLP	12	1.2	0.1
🟧 CLP_final	9	75.0	0.0

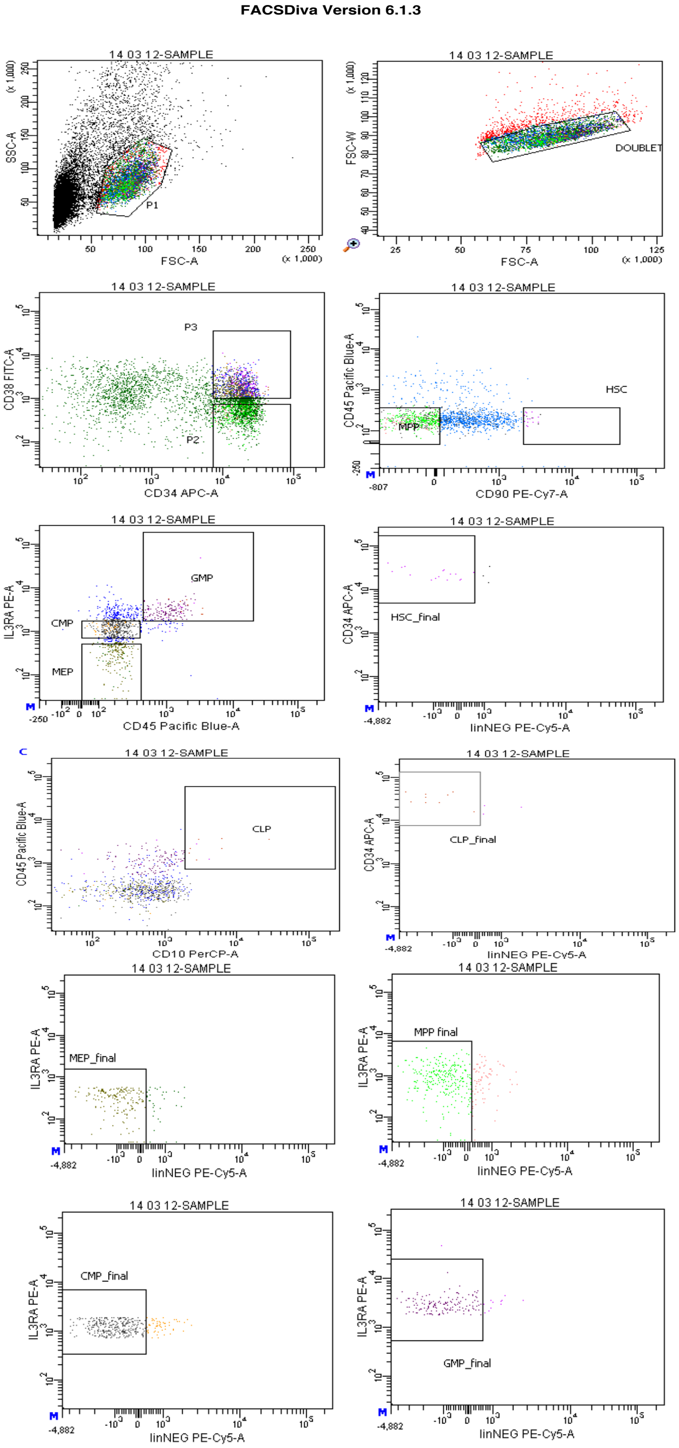


Figure S1: Example of progenitor cell sorting and gate setting. Gates were established using a fluorescence minus one strategy.

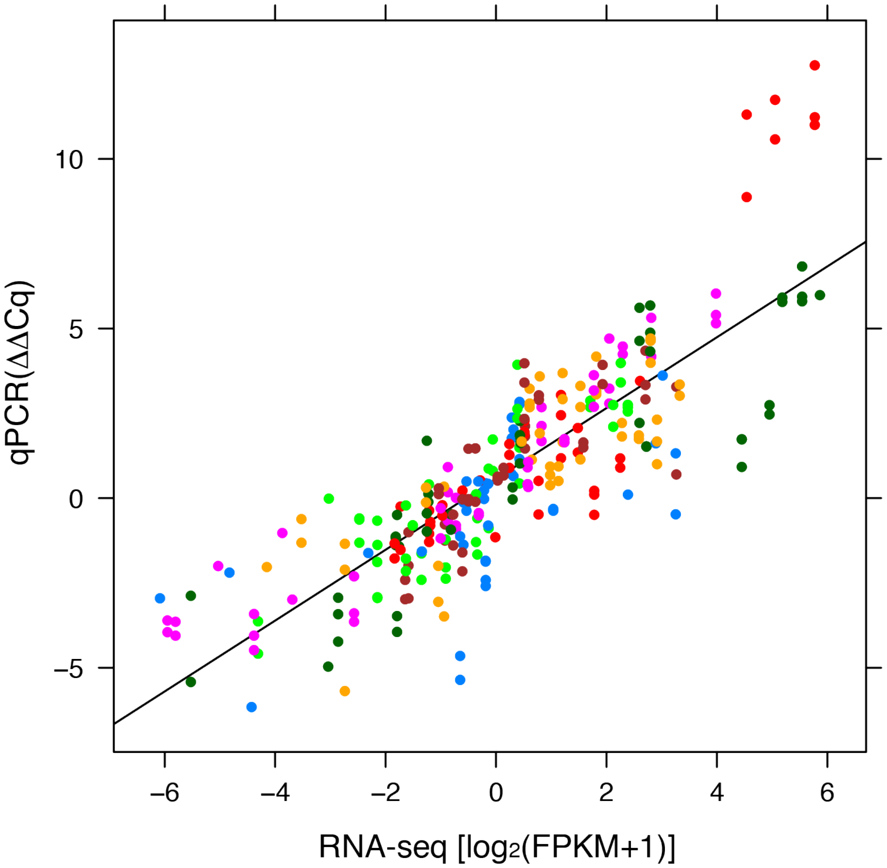


Figure S2: Correlation between the ΔΔCq of transcript-specific qPCR and quantification of transcript expression from MMSEQ analysis of RNA-seq data (R^2 = 0.70).

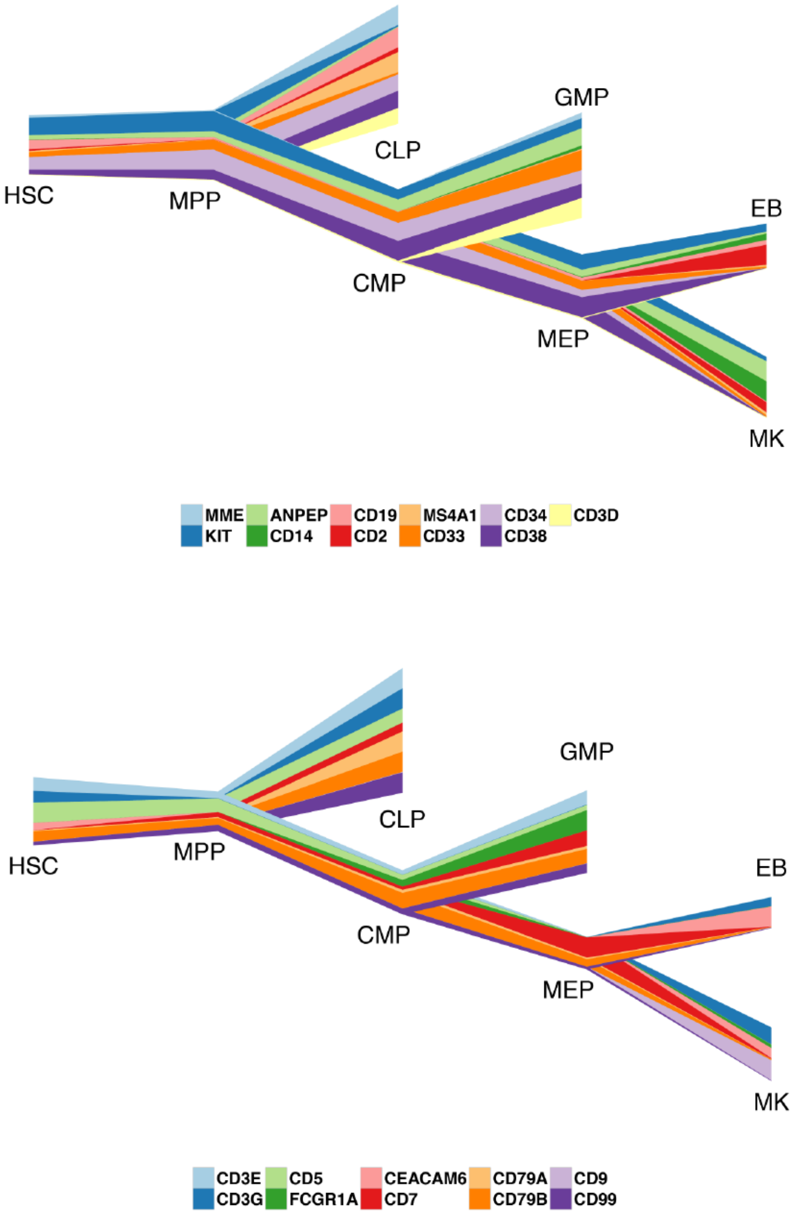


Figure S3: River plots illustrating gene expression levels for Cluster of Differentiation (CD) markers used in routine diagnosis of hematological malignancies (e.g. leukemia and lymphoma) across our dataset. Width corresponds to expression level relative to the maximum expression level of the gene across all cell types.

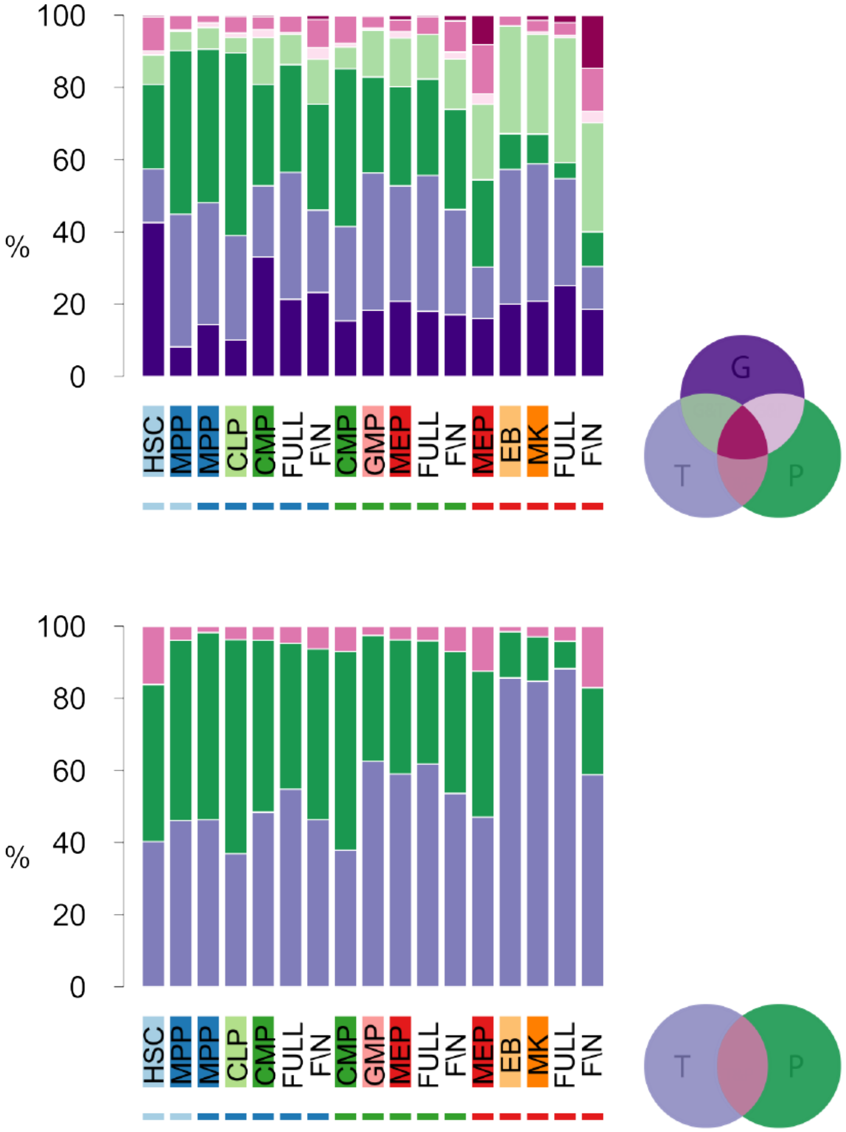
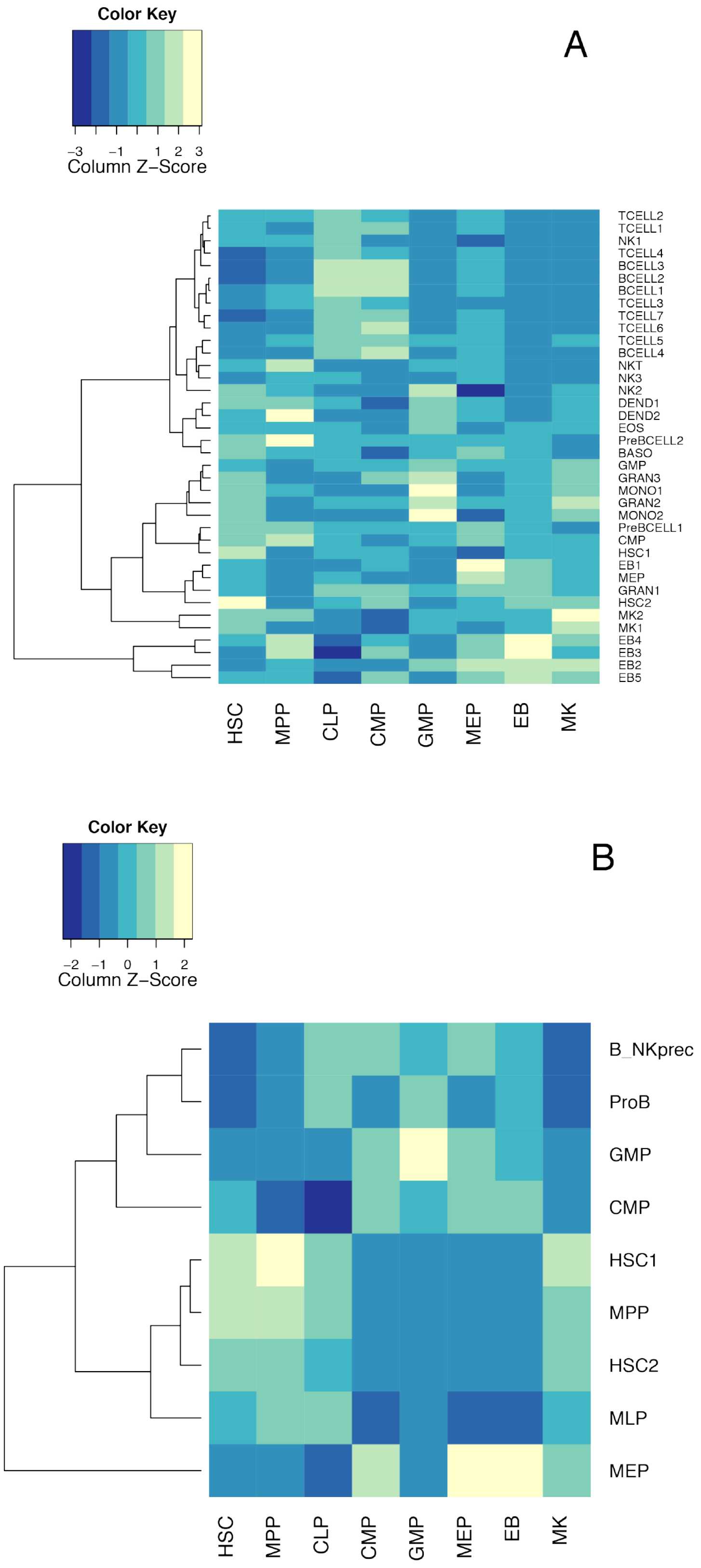
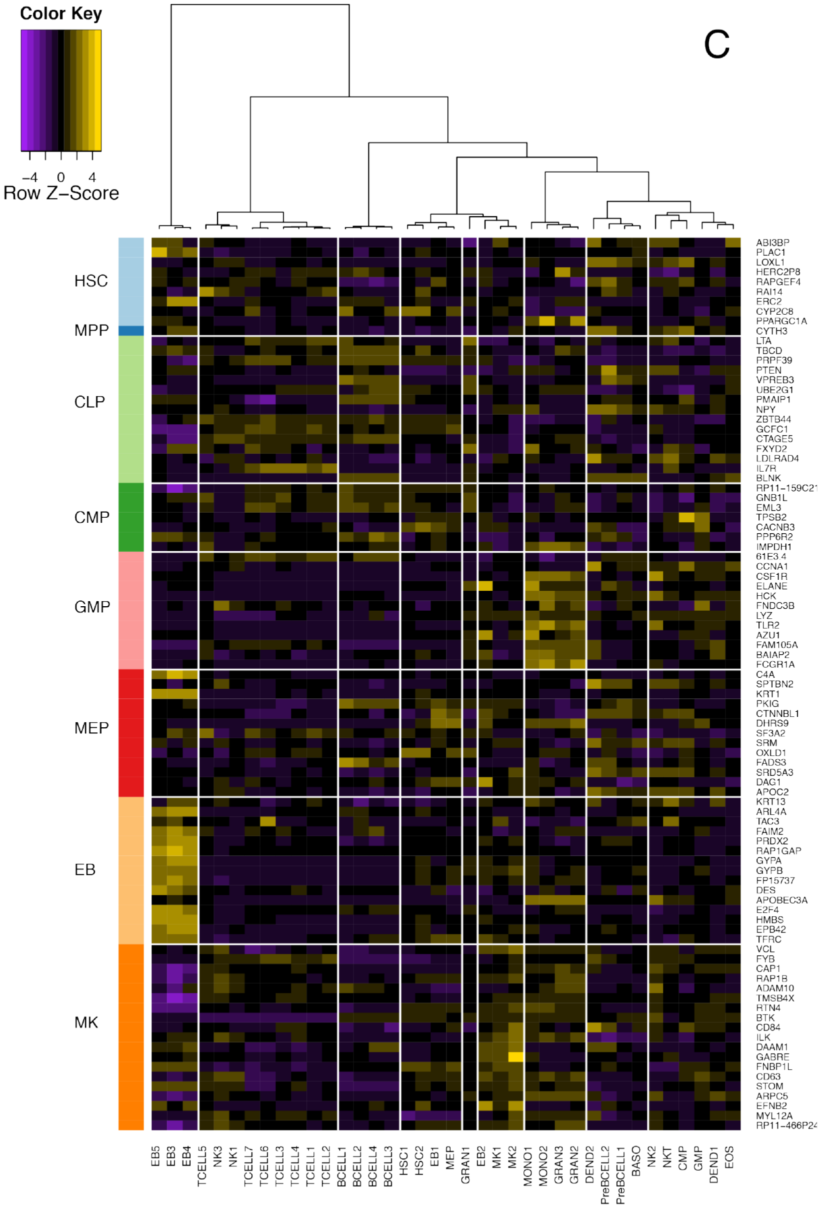


Figure S4: Overlap between sets of differentially expressed genes (G), transcripts (T) and transcript usage ratios (P), at either gene (upper panel) or transcript level (lower panel), at each lineage commitment point (lower bars in the x-axis), for the models used in the polytomous analysis (upper bars in the x-axis). Model F\N excludes the elements of model NULL from those in model FULL.



Color Key
−3
−1
1
2
3
Column Z−Score
A
TCELL2
TCELL1
NK1
TCELL4
BCELL3
BCELL2
BCELL1
TCELL3
TCELL7
TCELL6
TCELL5
BCELL4
NKT
NK3
NK2
DEND1
DEND2
EOS
PreBCELL2
BASO
GMP
GRAN3
MONO1
GRAN2
MONO2
PreBCELL1
CMP
HSC1
EB1
MEP
GRAN1
HSC2
MK2
MK1
EB4
EB3
EB2
EB5
HSC
MPP
CLP
CMP
GMP
MEP
EB
MK
Color Key
−2
−1
0
1
2
Column Z−Score
B
B_NKprec
ProB
GMP
CMP
HSC1
MPP
HSC2
MLP
MEP
HSC
MPP
CLP
CMP
GMP
MEP
EB
MK


Color Key
−4
0
4
Row Z−Score
C
HSC
MPP
CLP
CMP
GMP
MEP
EB
MK
ABI3BP
PLAC1
LOXL1
HERC2P8
RAPGEF4
RAI14
ERC2
CYP2C8
PPARGC1A
CYTH3
LTA
TBCD
PRPF39
PTEN
VPREB3
UBE2G1
PMAIP1
NPY
ZBTB44
GCFC1
CTAGE5
FXYD2
LDLRAD4
IL7R
BLNK
RP11−159C21
GNB1L
EML3
TPSB2
CACNB3
PPP6R2
IMPDH1
61E3.4
CCNA1
CSF1R
ELANE
HCK
FNDC3B
LYZ
TLR2
AZU1
FAM105A
BAIAP2
FCGR1A
C4A
SPTBN2
KRT1
PKIG
CTNNBL1
DHRS9
SF3A2
SRM
OXLD1
FADS3
SRD5A3
DAG1
APOC2
KRT13
ARL4A
TAC3
FAIM2
PRDX2
RAP1GAP
GYPA
GYPB
FP15737
DES
APOBEC3A
E2F4
HMBS
EPB42
TFRC
VCL
FYB
CAP1
RAP1B
ADAM10
TMSB4X
RTN4
BTK
CD84
ILK
DAAM1
GABRE
FNBP1L
CD63
STOM
ARPC5
EFNB2
MYL12A
RP11−466P24
EB5
EB3
EB4
TCELL5
NK3
NK1
TCELL7
TCELL6
TCELL3
TCELL4
TCELL1
TCELL2
BCELL1
BCELL2
BCELL4
BCELL3
HSC1
HSC2
EB1
MEP
GRAN1
EB2
MK1
MK2
MONO1
MONO2
GRAN3
GRAN2
DEND2
PreBCELL2
PreBCELL1
BASO
NK2
NKT
CMP
GMP
DEND1
EOS

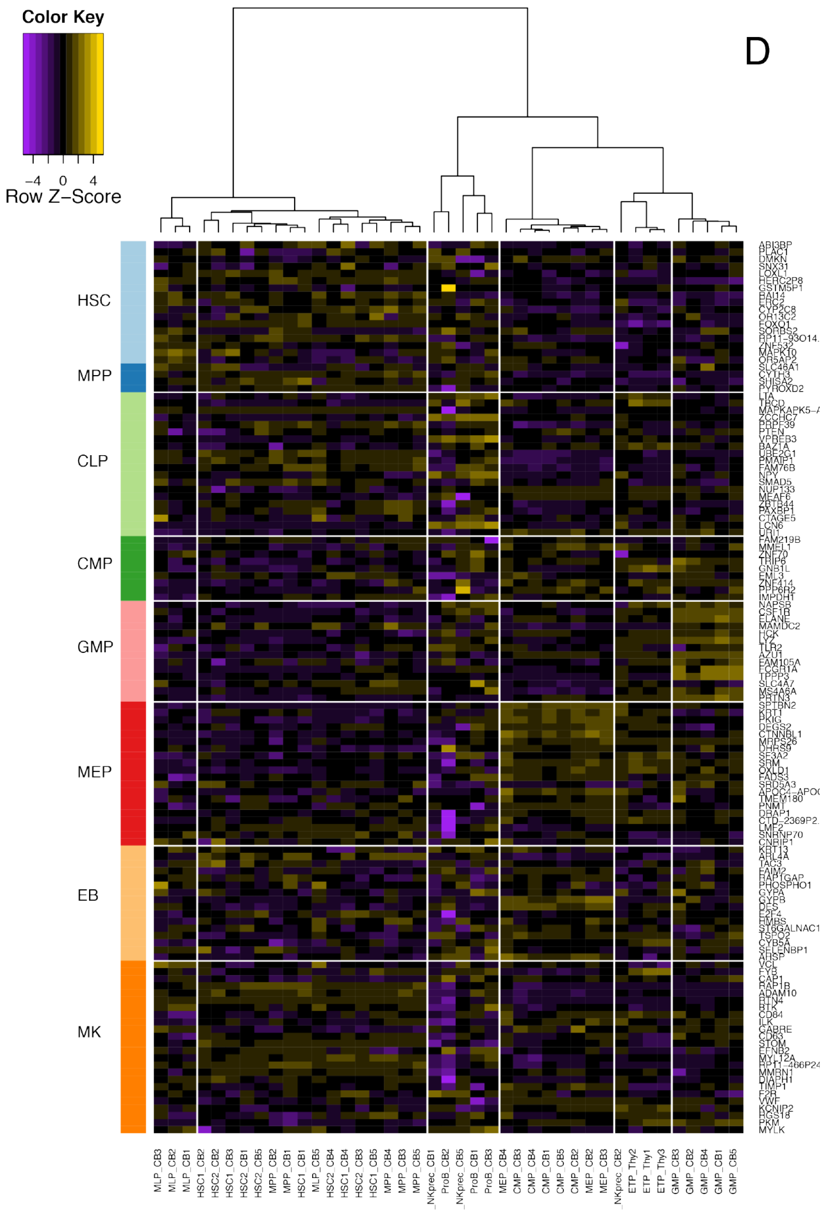


Figure S5: Comparison of our data with previous microarray studies based on lineage-specific genes. Using the test statistic from the Wilcoxon rank sum test that the lineage-specific gene set is more highly expressed than all other genes expressed in the study: **(A)** microarray data from Novernshtern *et al.* (14); and **(B)** from de Laurenti *et al.* (16). **(C)** Heatmap of lineage-specific genes identified by this study visualized using previously published microarray-based data from the compendium of gene expression in hematopoiesis in Novernshtern *et al.* where biological replicates have been summarized (14). **(D)** Heatmap of lineage-specific genes identified by this study visualized using previously published microarray-based data from the compendium of gene expression in hematopoiesis in Laurenti *et al.* with all biological replicates shown (16).

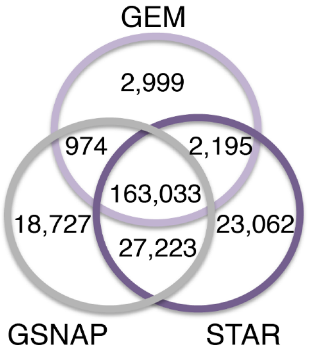


Figure S6: Number of splice junctions identified by GEM, GSNAP and STAR for all cell types.

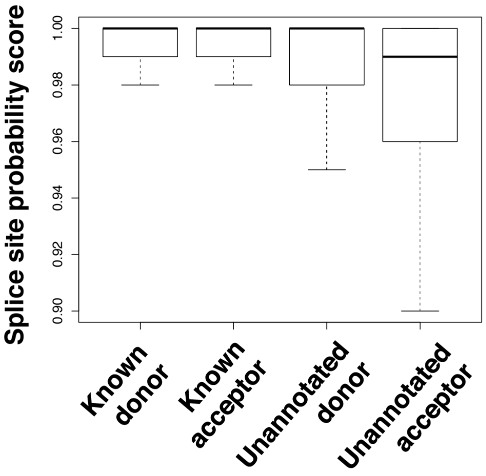


Figure S7: Splice site probability scores for known and unannotated splice junctions.

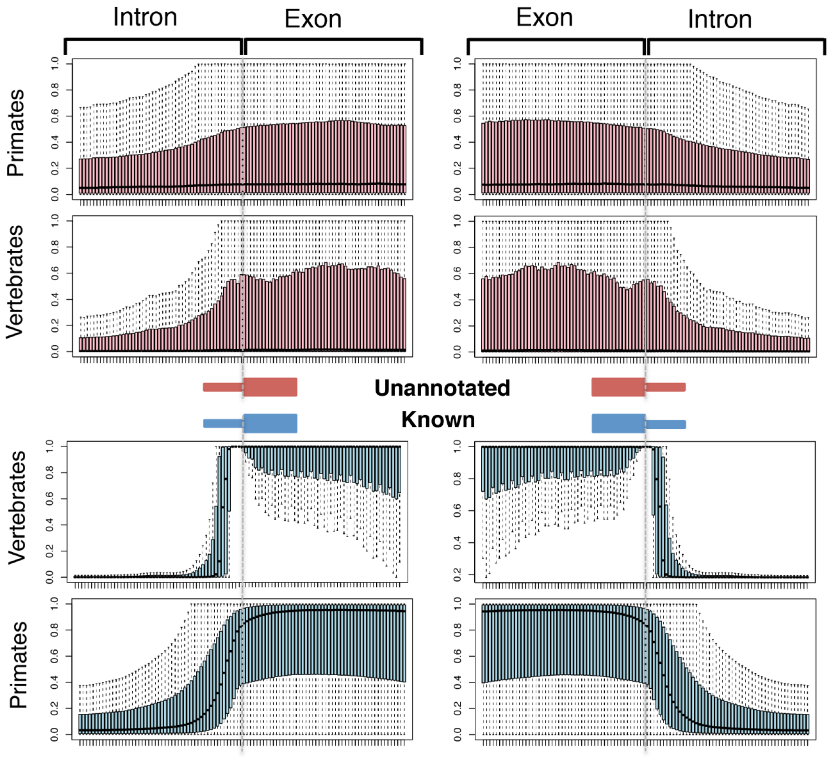


Figure S8: PhastCons conservation scores in a 100 bp window around known (blue) and unannotated (red) splice junctions. The search window includes 50 bp of exonic and 50 bp of intronic regions around the splice junctions in vertebrates and primates.

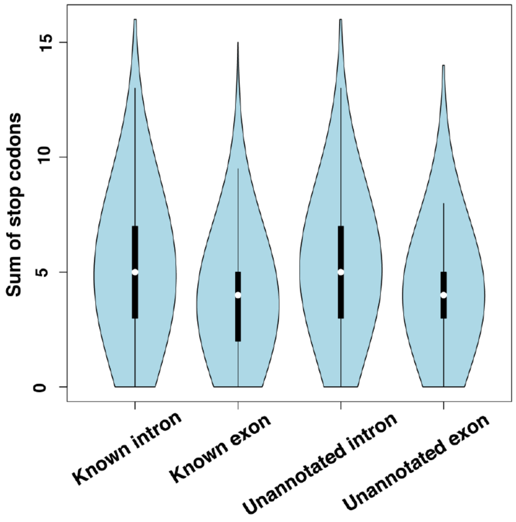


Figure S9: Distribution of the sum of stop codons around known and unannotated splice junctions in all six open reading frames.

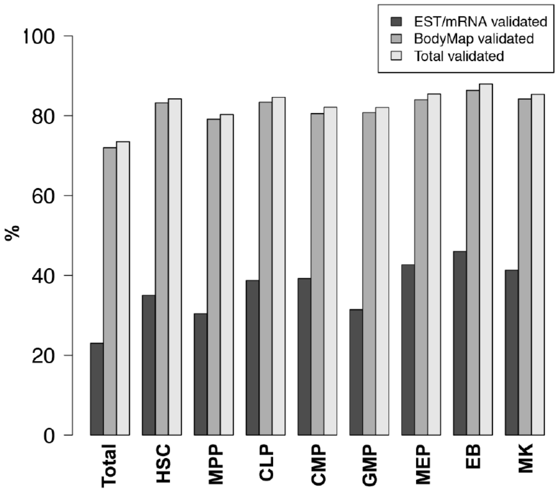


Figure S10: Percentage of unannotated splice junctions validated in the EST/mRNA, the BodyMap 2.0, or both datasets ("Total validated").

%	GT-AG	GC-AG	AT-AC	Other
Known	98.0	0.9	0.1	1.0
Unannotated	95.0	2.2	0.0	2.8
Novel	90.7	7.3	0.0	2.0

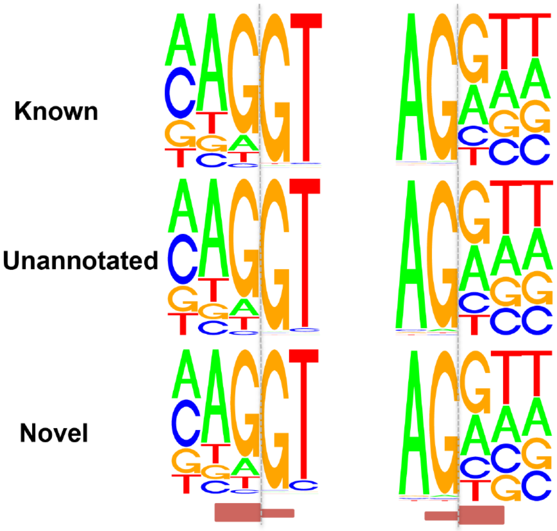


Figure S11: Splice site motifs for annotated, unannotated and novel donors and acceptors.

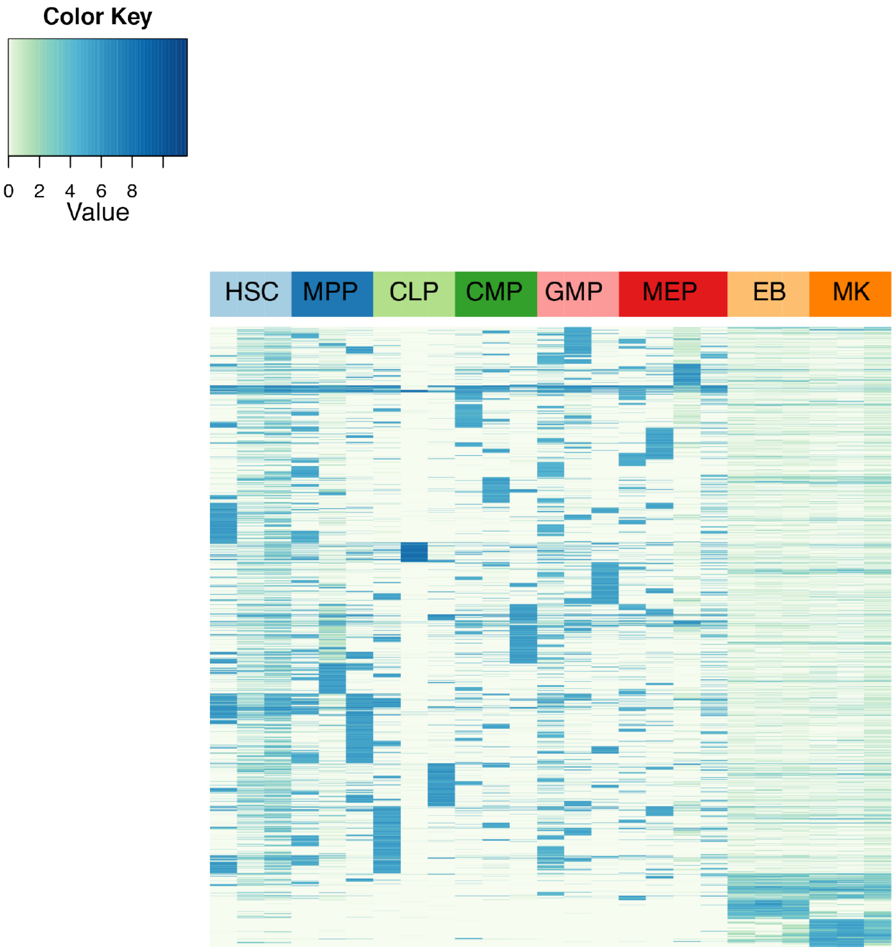
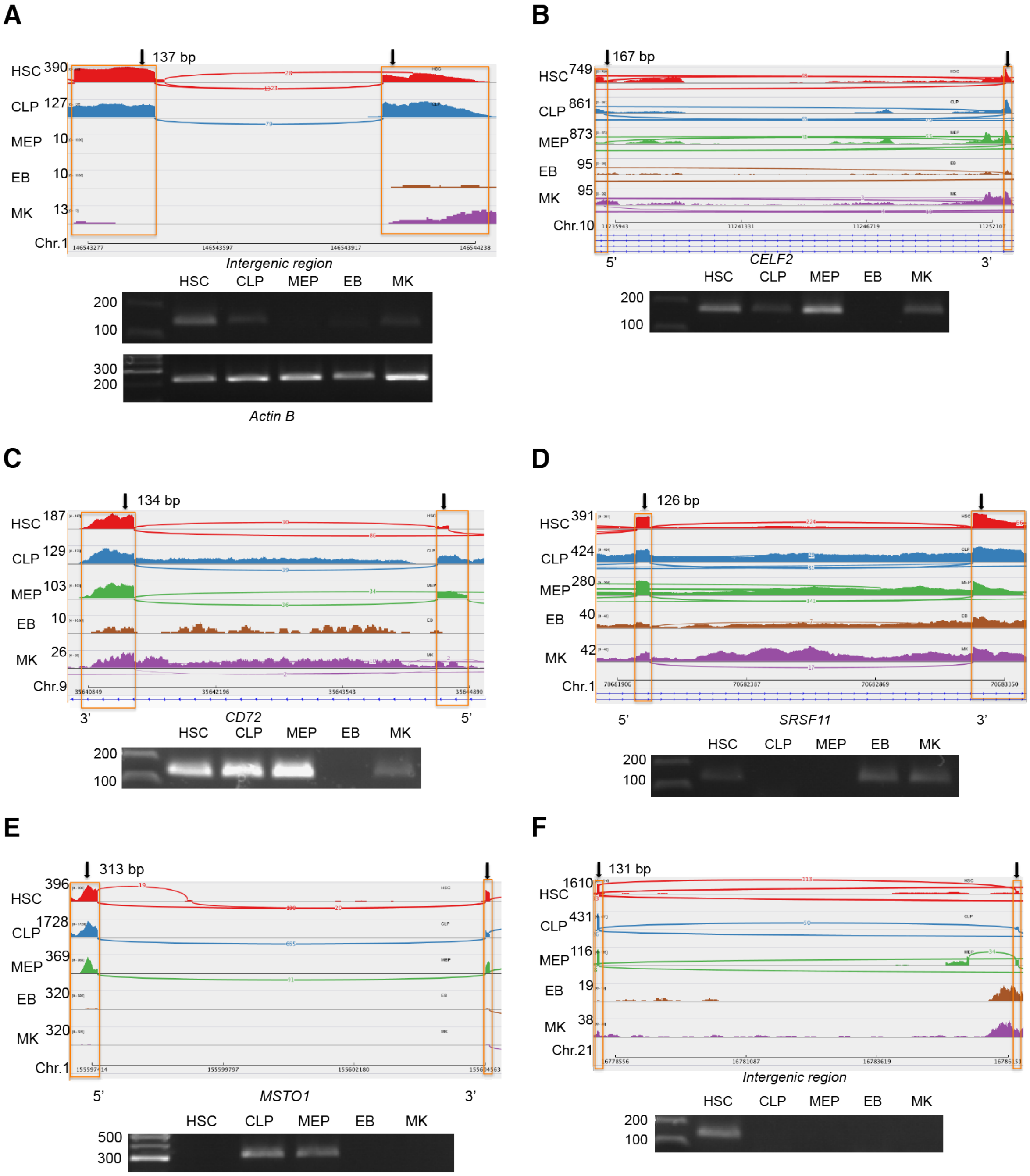
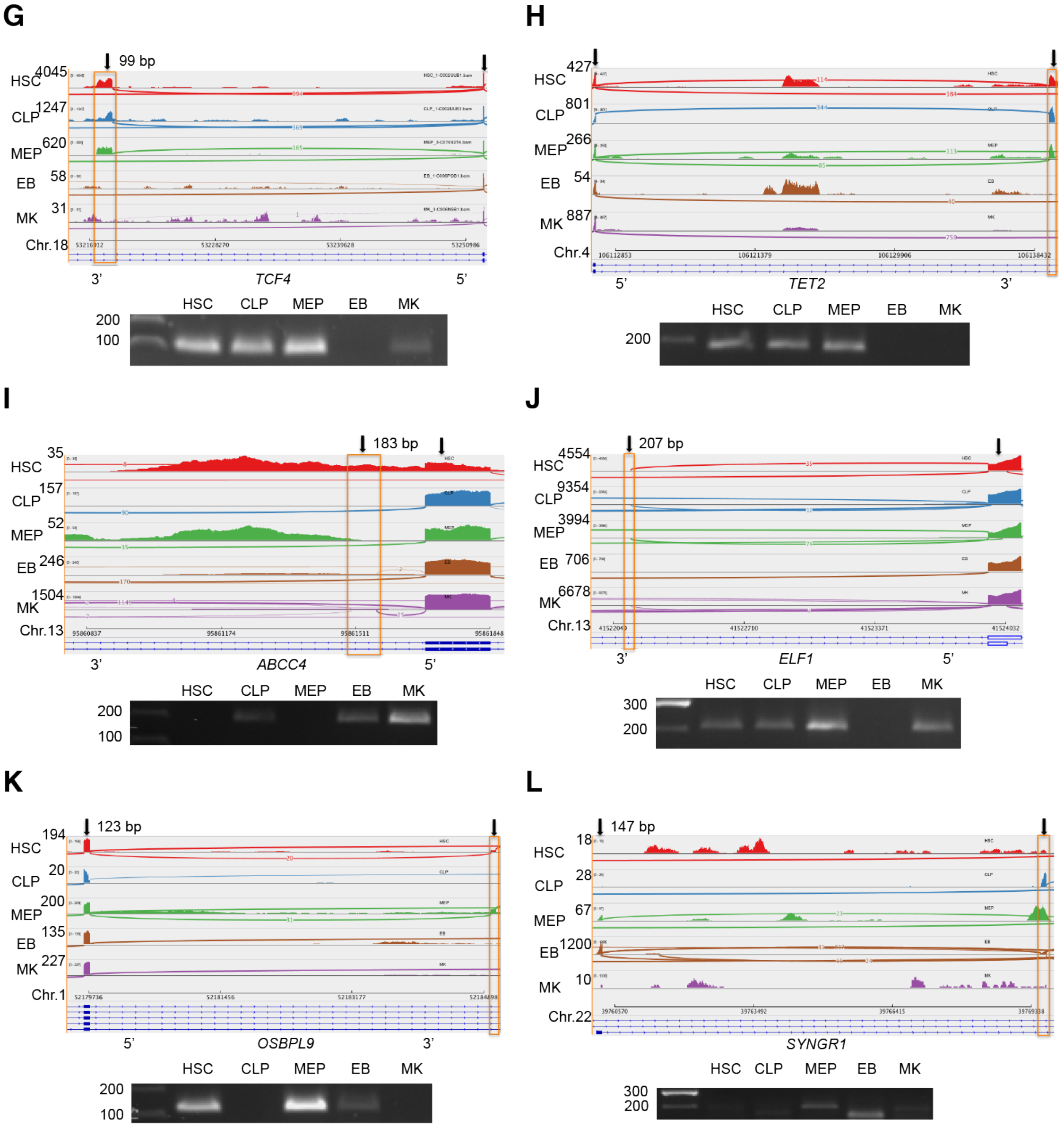


Figure S12: Normalized read counts of novel splice junctions in progenitors. Normalization was based on sample size.


A
137 bp
HSC 390
CLP 127
MEP 10
EB 10
MK 13
Chr.1
Intergenic region
HSC CLP MEP EB MK
200
100
300
200
Actin B
B
167 bp
HSC 749
CLP 861
MEP 873
EB 95
MK 95
Chr.10
5'
CELF2
3'
HSC CLP MEP EB MK
200
100
C
134 bp
HSC 187
CLP 129
MEP 103
EB 10
MK 26
Chr.9
3'
CD72
5'
HSC CLP MEP EB MK
200
100
D
126 bp
HSC 391
CLP 424
MEP 280
EB 40
MK 42
Chr.1
5'
SRSF11
3'
HSC CLP MEP EB MK
200
100
E
313 bp
HSC 396
CLP 1728
MEP 369
EB 320
MK 320
Chr.1
5'
MSTO1
3'
HSC CLP MEP EB MK
500
300
F
131 bp
HSC 1610
CLP 431
MEP 116
EB 19
MK 38
Chr.21
Intergenic region
HSC CLP MEP EB MK
200
100


G
99 bp
HSC 4045
CLP 1247
MEP 620
EB 58
MK 31
Chr.18
53216912
53228270
53239628
53250986
3'
TCF4
5'
HSC CLP MEP EB MK
200
100
H
HSC 427
CLP 801
MEP 266
EB 54
MK 887
Chr.4
106112853
106121379
106129906
106138432
5'
TET2
3'
HSC CLP MEP EB MK
200
I
183 bp
HSC 35
CLP 157
MEP 52
EB 246
MK 1504
Chr.13
95860837
95861174
95861511
95861848
3'
ABCC4
5'
HSC CLP MEP EB MK
200
100
J
207 bp
HSC 4554
CLP 9354
MEP 3994
EB 706
MK 6678
Chr.13
41522049
41522710
41523371
41524032
3'
ELF1
5'
HSC CLP MEP EB MK
300
200
K
123 bp
HSC 194
CLP 20
MEP 200
EB 135
MK 227
Chr.1
52179736
52181456
52183177
52184898
5'
OSBPL9
3'
HSC CLP MEP EB MK
200
100
L
147 bp
HSC 18
CLP 28
MEP 67
EB 1200
MK 10
Chr.22
39760570
39763492
39766415
39769338
SYNGR1
HSC CLP MEP EB MK
300
200

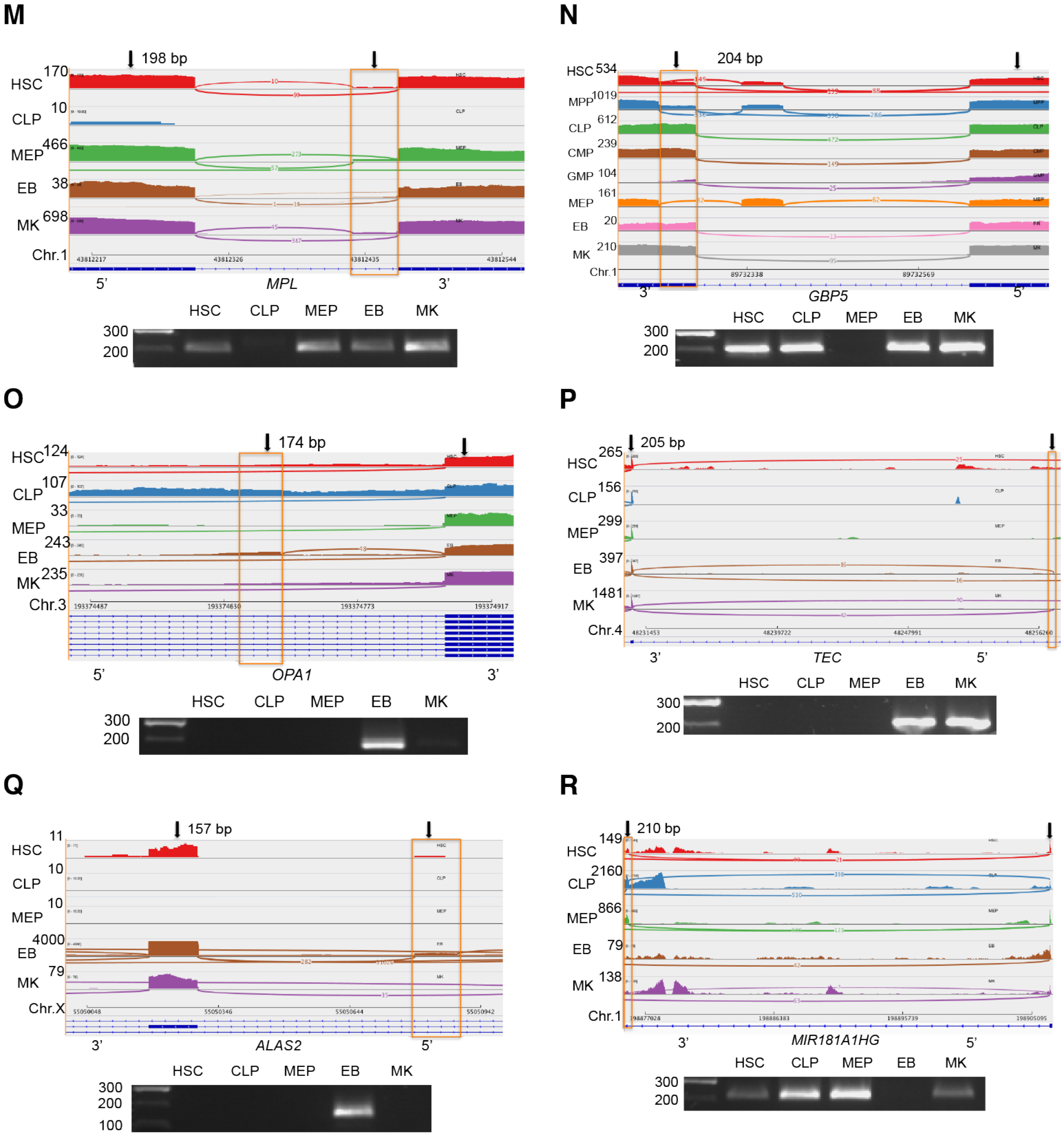


Figure S13: PCR validation of novel splice junctions. For each novel splice junction, the sashimi plot of RNA-seq reads in HSCs, CLPs, MEPs, EBs and MKs are shown in the top panel. The lines connecting two exons in the sashimi plot represent the novel splice junction with the number of reads supporting it within the line. The novel exons are boxed in orange. The two black arrows on the top of the sashimi plot indicate the positions of primers with the estimated length of PCR product next to the left primer. PCR products visualized on gel are in the lower panel. Size markers are in base pairs. The lower panel in panel **A** displays the PCR fragment for Actin B, used as a positive control for each sample.

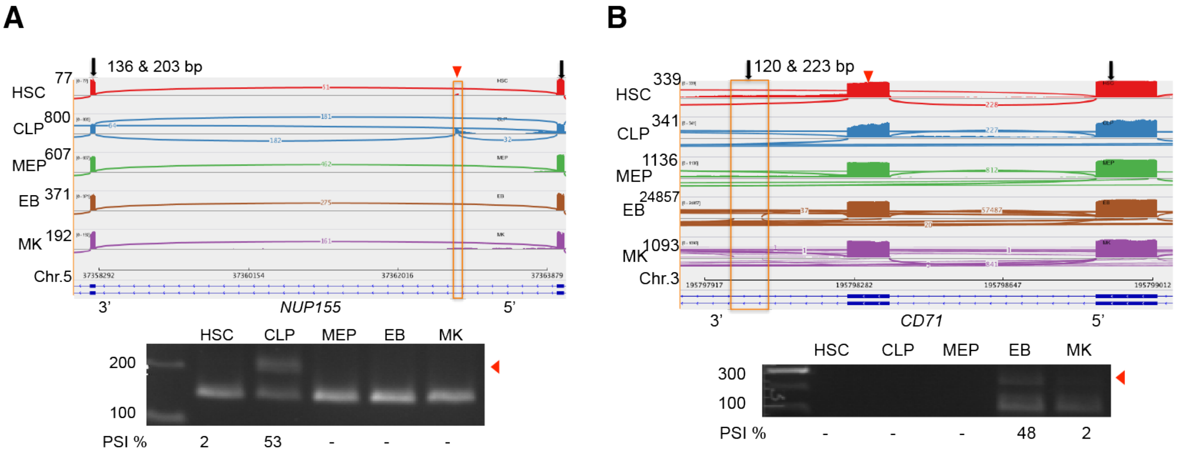


Figure S14: Example of PCR validation of novel splice junctions with DSU. For each novel splice junction, the sashimi plot of RNA-seq reads in HSCs, CLPs, MEPs, EBs and MKs are shown in the top panel. The lines connecting two exons in the sashimi plot represent the novel splice junction with the number of reads supporting it within the line. The novel exons are boxed in orange. The two black arrows on the top of the sashimi plot indicate the positions of primers with the estimated length of PCR product next to the left primer. PCR products visualized on gel are in the lower panel. Size markers are in base pairs. Red arrows indicate the PCR product that contains the DSU exon and the position of this DSU in the sashimi plot. Values for Percentage of Splicing In (PSI) for each cell type with DSU were computed from three biological replicates and displayed under the appropriate PCR fragment.

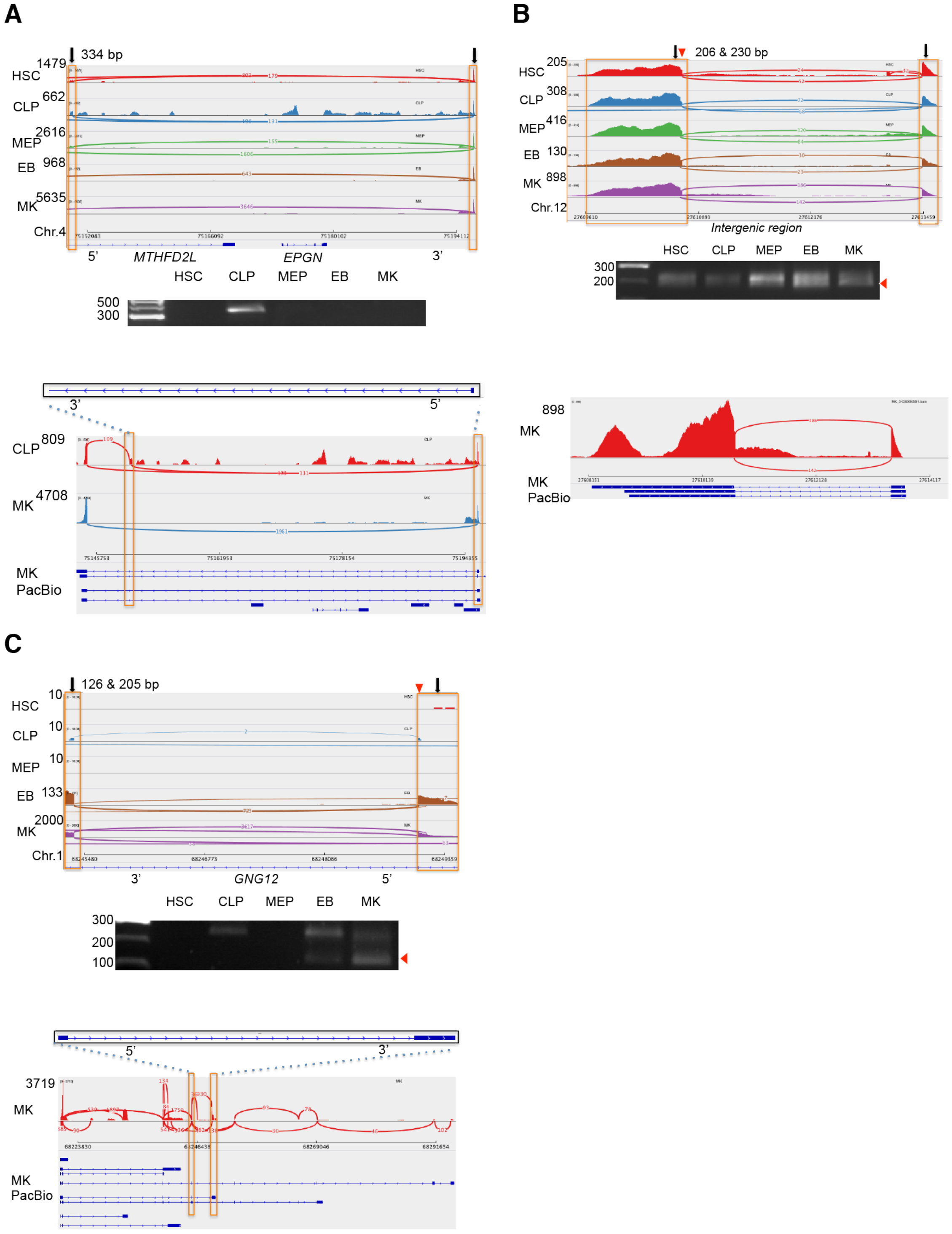
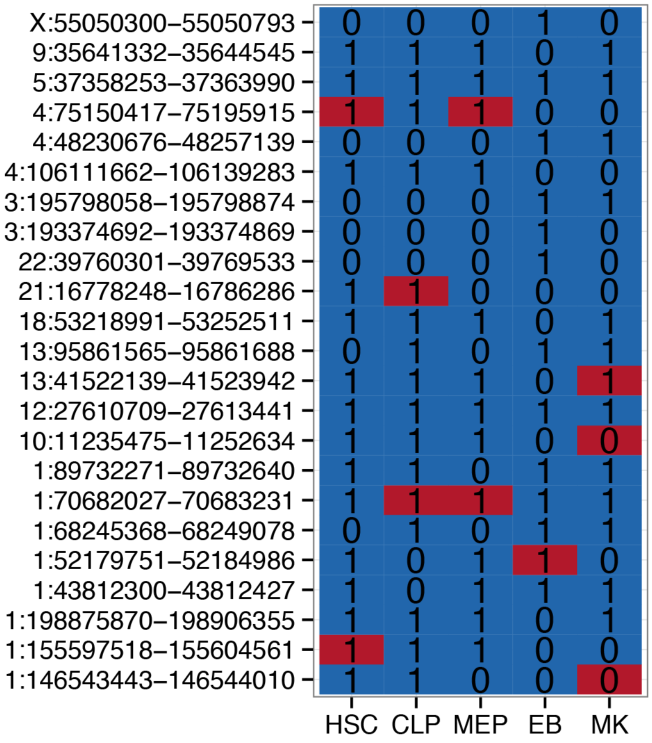


Figure S15: PCR validation for novel splice junctions supported by full-length transcripts. For each novel splice junction, the sashimi plot of RNA-seq reads in HSCs, CLPs, MEPs, EBs and MKs are shown in the top panel. The lines connecting two exons in the sashimi plot represent the novel splice junction with the number of reads supporting it within the line. The novel exons are boxed in orange. The two black arrows on the top of the sashimi plot indicate the positions of primers with the estimated length of PCR product next to the left primer. PCR products visualized on gel are in the lower panel. Size markers are in base pairs. Red arrows indicate the PCR product that contains the DSU exon and the position of this alternative splice junction within the sashimi plot. In the lower panel, we display the RNA-seq reads from one MK sample and the reads from the MK PacBio data.



	HSC	CLP	MEP	EB	MK
X:55050300–55050793	0	0	0	1	0
9:35641332–35644545	1	1	1	0	1
5:37358253–37363990	1	1	1	1	1
4:75150417–75195915	🟥 1	1	🟥 1	0	0
4:48230676–48257139	0	0	0	1	1
4:106111662–106139283	1	1	1	0	0
3:195798058–195798874	0	0	0	1	1
3:193374692–193374869	0	0	0	1	0
22:39760301–39769533	0	0	0	1	0
21:16778248–16786286	1	🟥 1	0	0	0
18:53218991–53252511	1	1	1	0	1
13:95861565–95861688	0	1	0	1	1
13:41522139–41523942	1	1	1	0	🟥 1
12:27610709–27613441	1	1	1	1	1
10:11235475–11252634	1	1	1	0	🟥 0
1:89732271–89732640	1	1	0	1	1
1:70682027–70683231	1	🟥 1	🟥 1	1	1
1:68245368–68249078	0	1	0	1	1
1:52179751–52184986	1	0	1	🟥 1	0
1:43812300–43812427	1	0	1	1	1
1:198875870–198906355	1	1	1	0	1
1:155597518–155604561	🟥 1	1	1	0	0
1:146543443–146544010	1	1	0	0	🟥 0

Figure S16: Concordance matrix of the validation of novel splice junctions. A blue box indicates concordance, whereas a red box indicates discordance between RNA-seq and PCR data. The number in each cell indicates if the splice junction is present (1) or absent (0) from the RNA-seq dataset. Each row label represents the chromosomal coordinates of the novel splice junction.

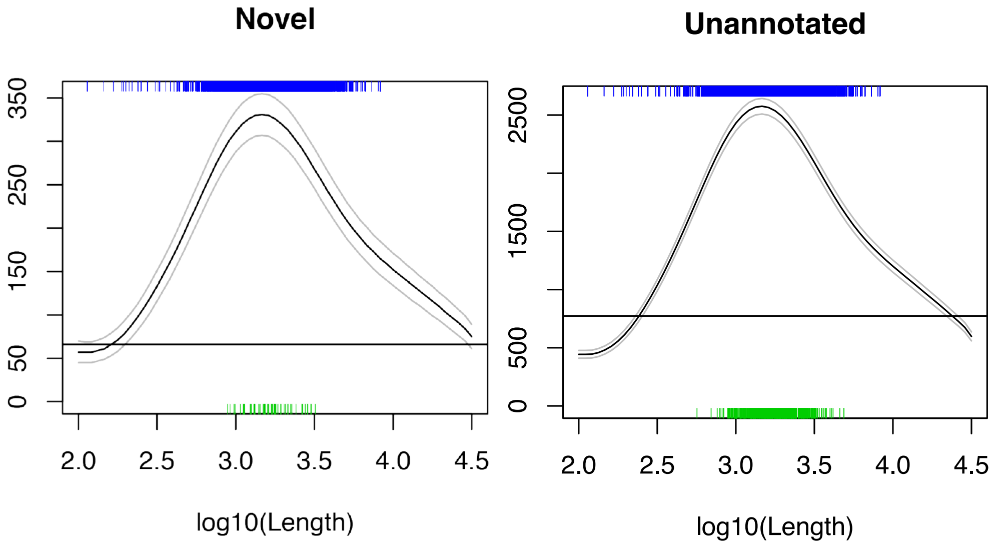


Figure S17: Number of splice junctions discovered by RNA-seq that should have been observed in PacBio as a function of length of the transcript containing that junction after adjusting for read depth. The blue lines at the top represent the junctions observed by both technologies and which were annotated in Ensembl. The green lines represent the junctions observed by both technologies and that were ”novel” (left) or ”unannotated” (right). The 5% and 95% confidence intervals are depicted. Note that validated junctions (green lines) were observed precisely where expected from the training data (Ensembl).

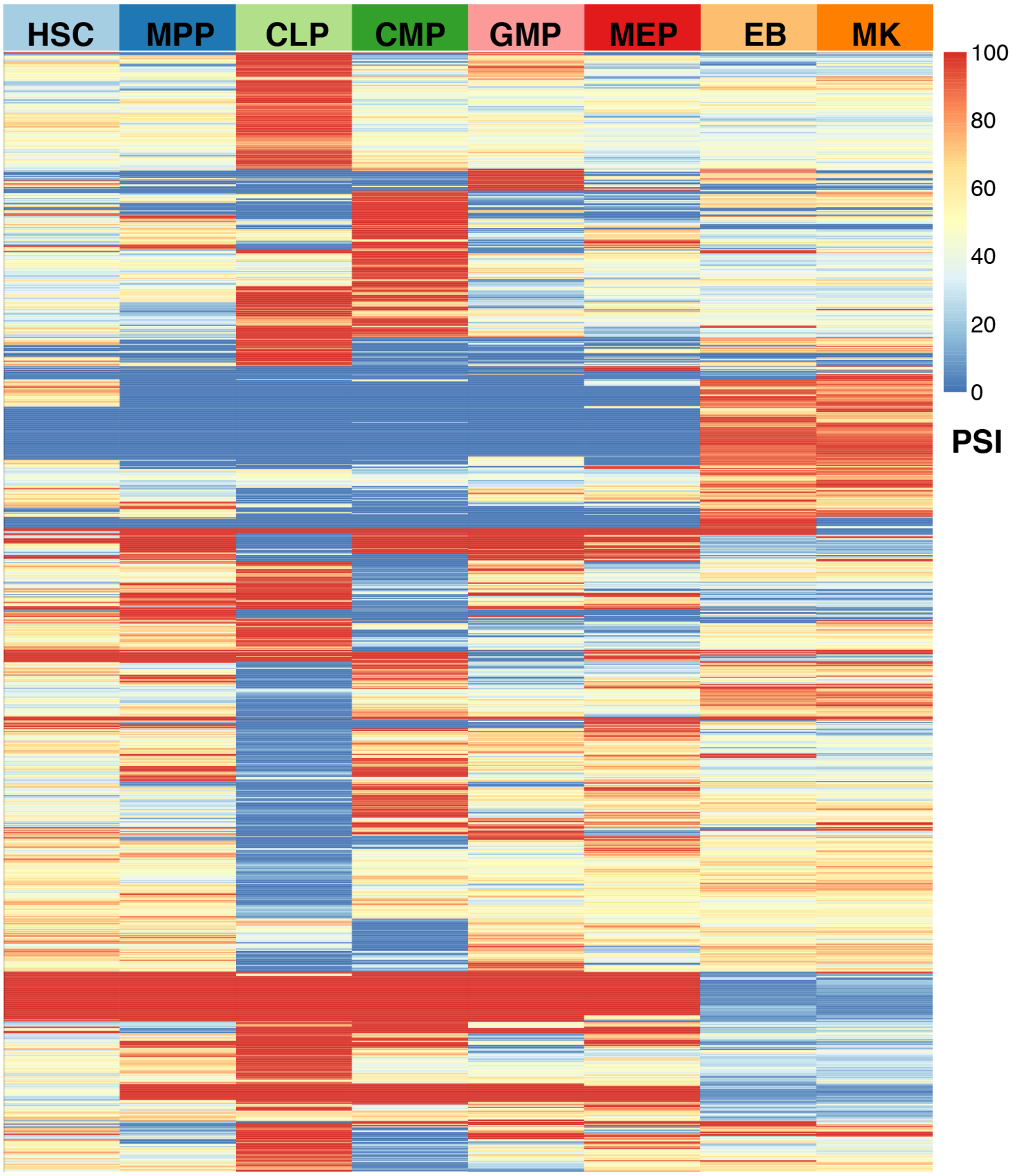


Figure S18: Percent-spliced-in (PSI) values in DSU splice junctions in the eight progenitors.

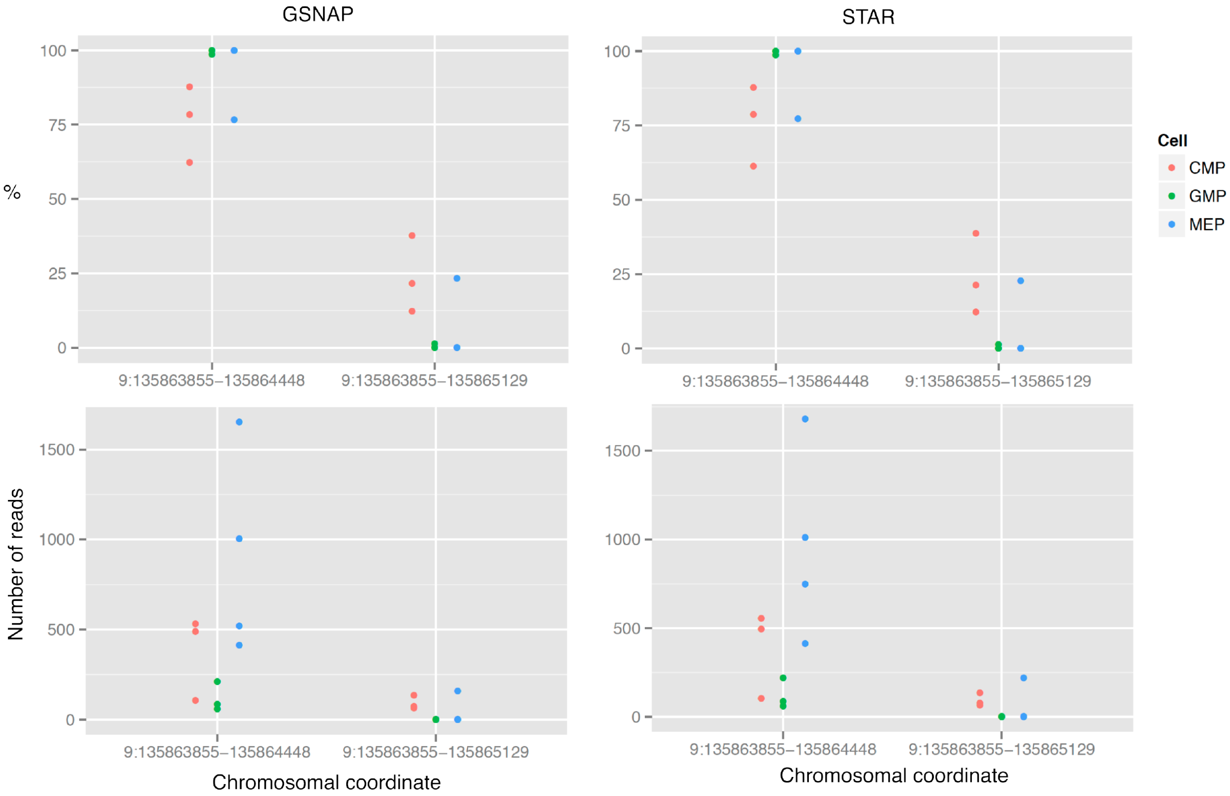


Figure S19: Differential splice junction usage in *GFI1B*. Upper panels show the percentage and lower panels the read counts, for both splice junctions using GSNAP (left) and STAR (right).

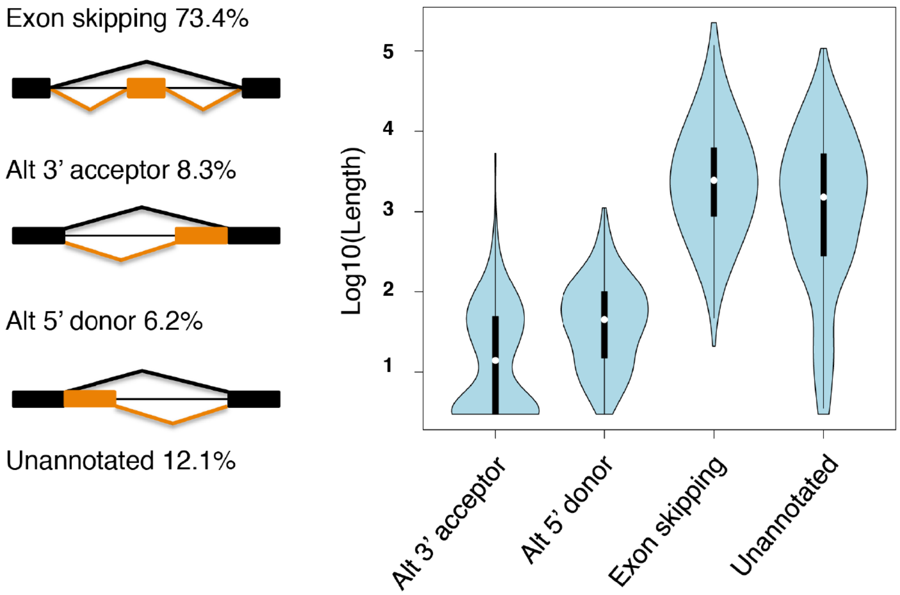


Figure S20: Types of alternative splicing with differential usage in hematopoiesis. Ensembl annotations were used as reference to define the different types of splicing events: exon skipping, alternative (alt) 3' acceptor, alternative 5' donor and unannotated. The left panel shows the frequency of these types of splicing events. The right panel shows the distribution of intron length for each of the alternative splicing events.

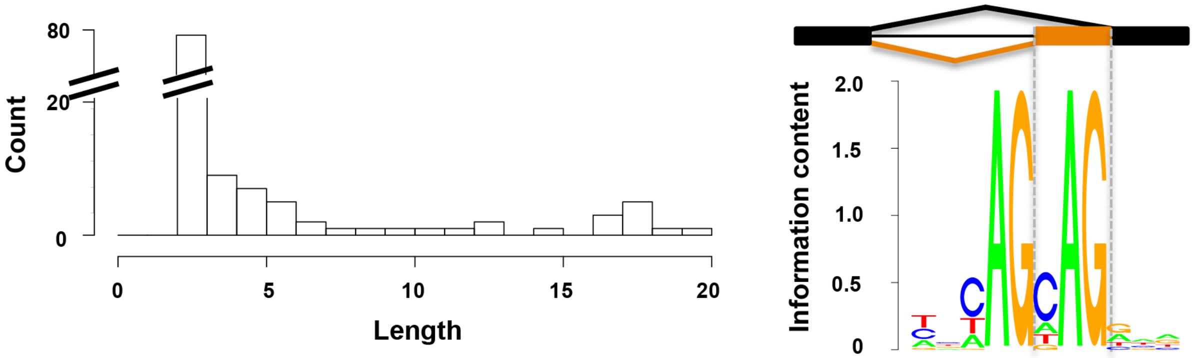


Figure S21: Distribution of the number of nucleotides between alternative splice junctions with DSU. Only splice junctions with two alternative 3'acceptor sites were considered. There was a high occurrence of a 3 bp nucleotide difference. The sequence underlying these 3 bp nucleotide regions has a NAGNAG motif, allowing for the addition of one amino acid, while maintaining the open reading frame.

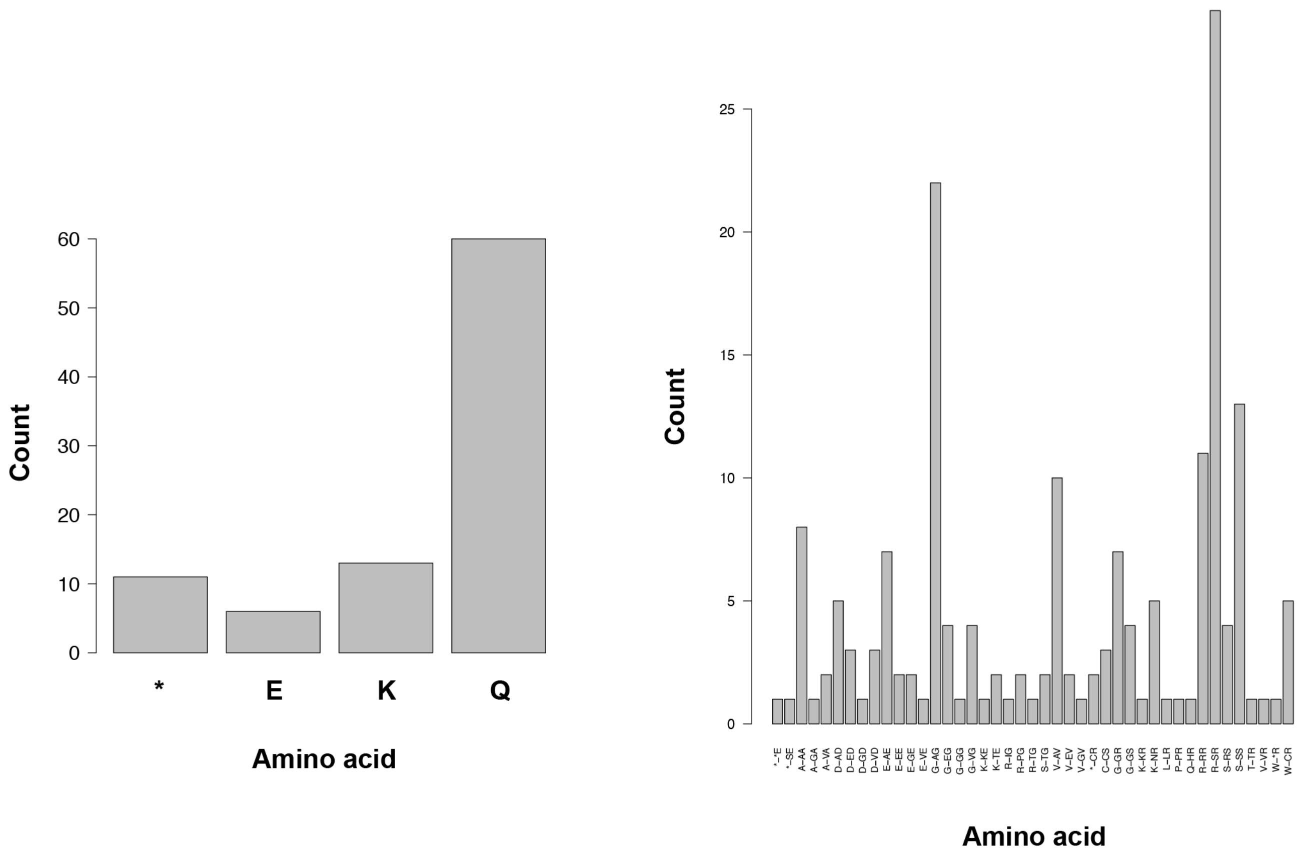
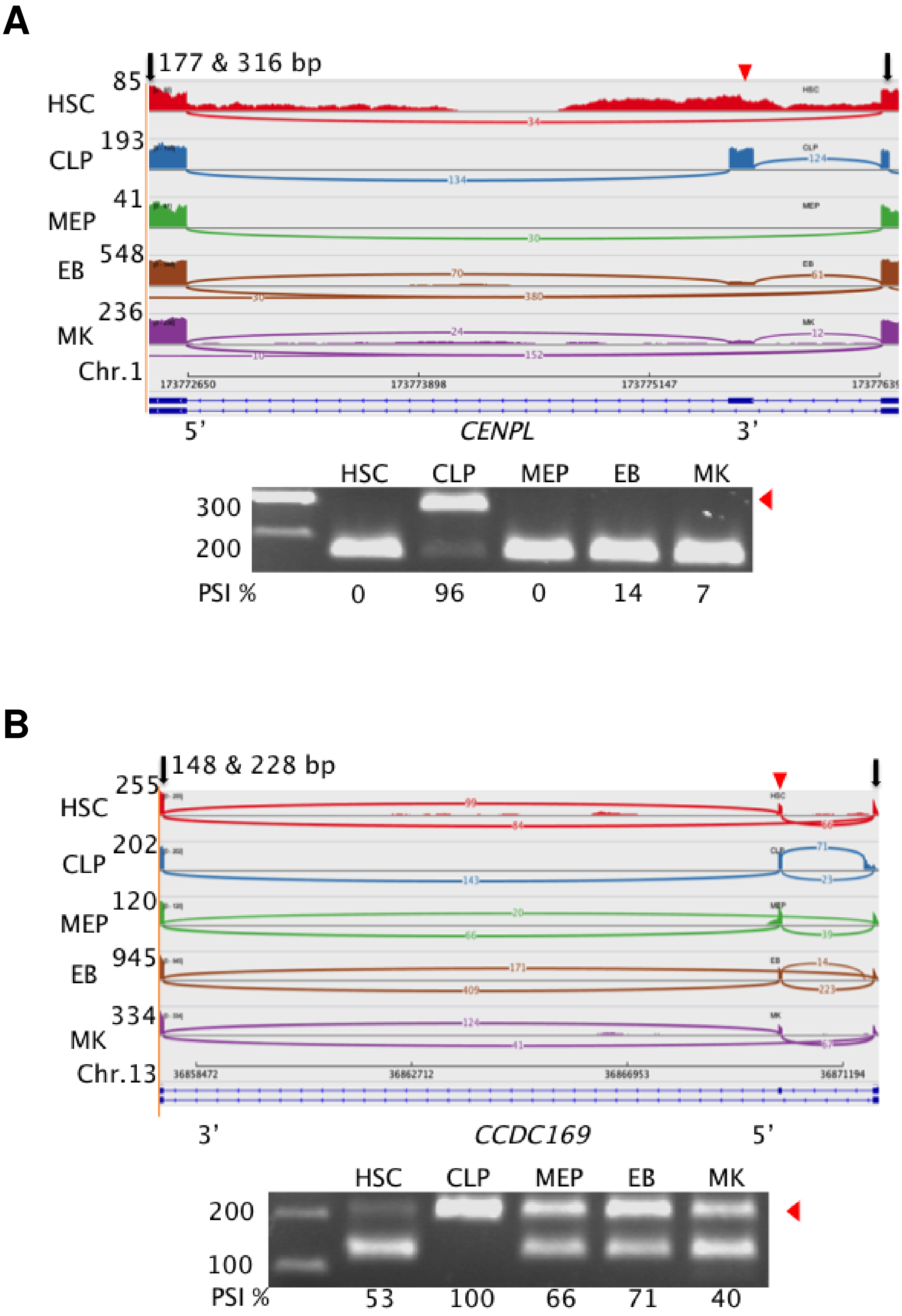
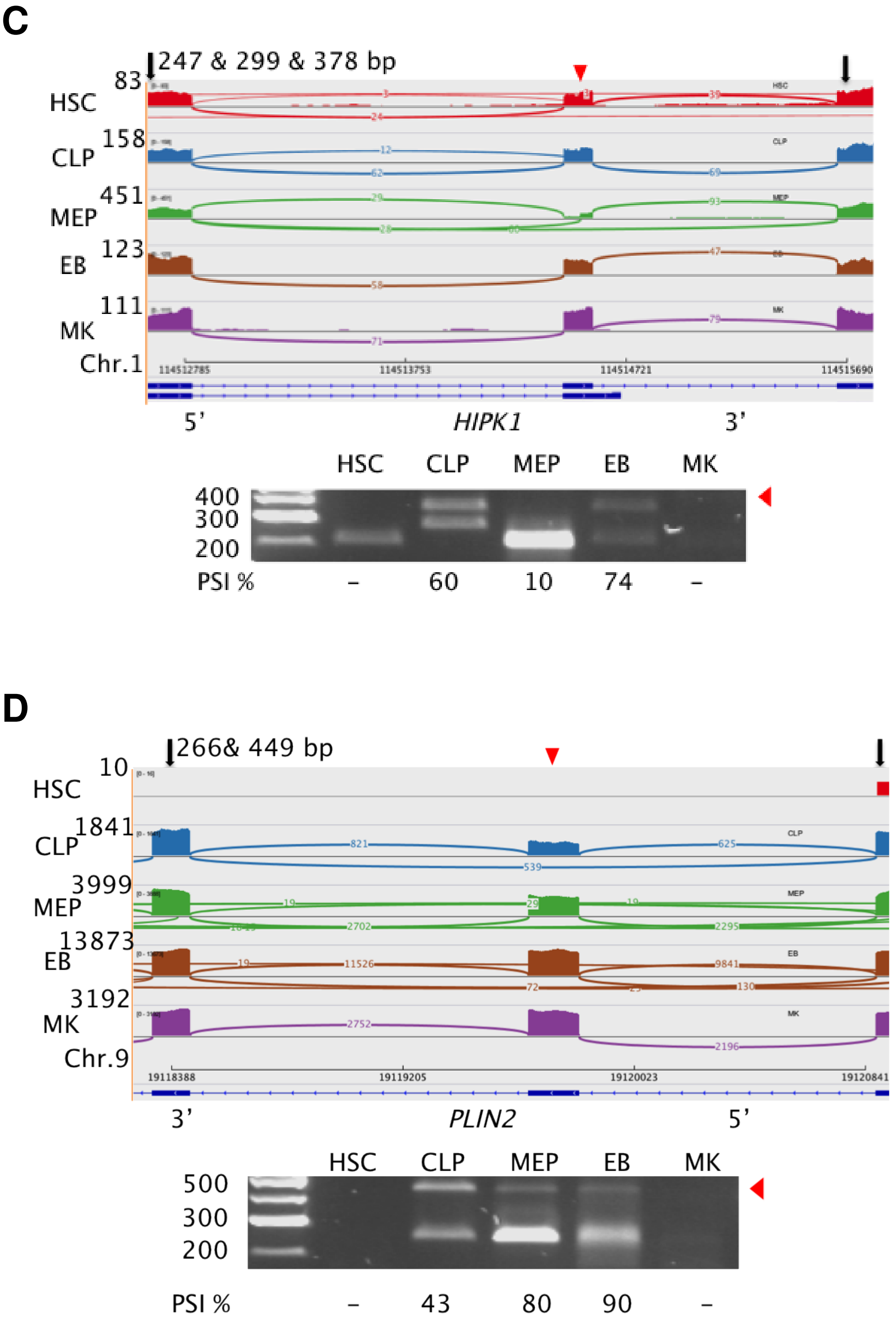
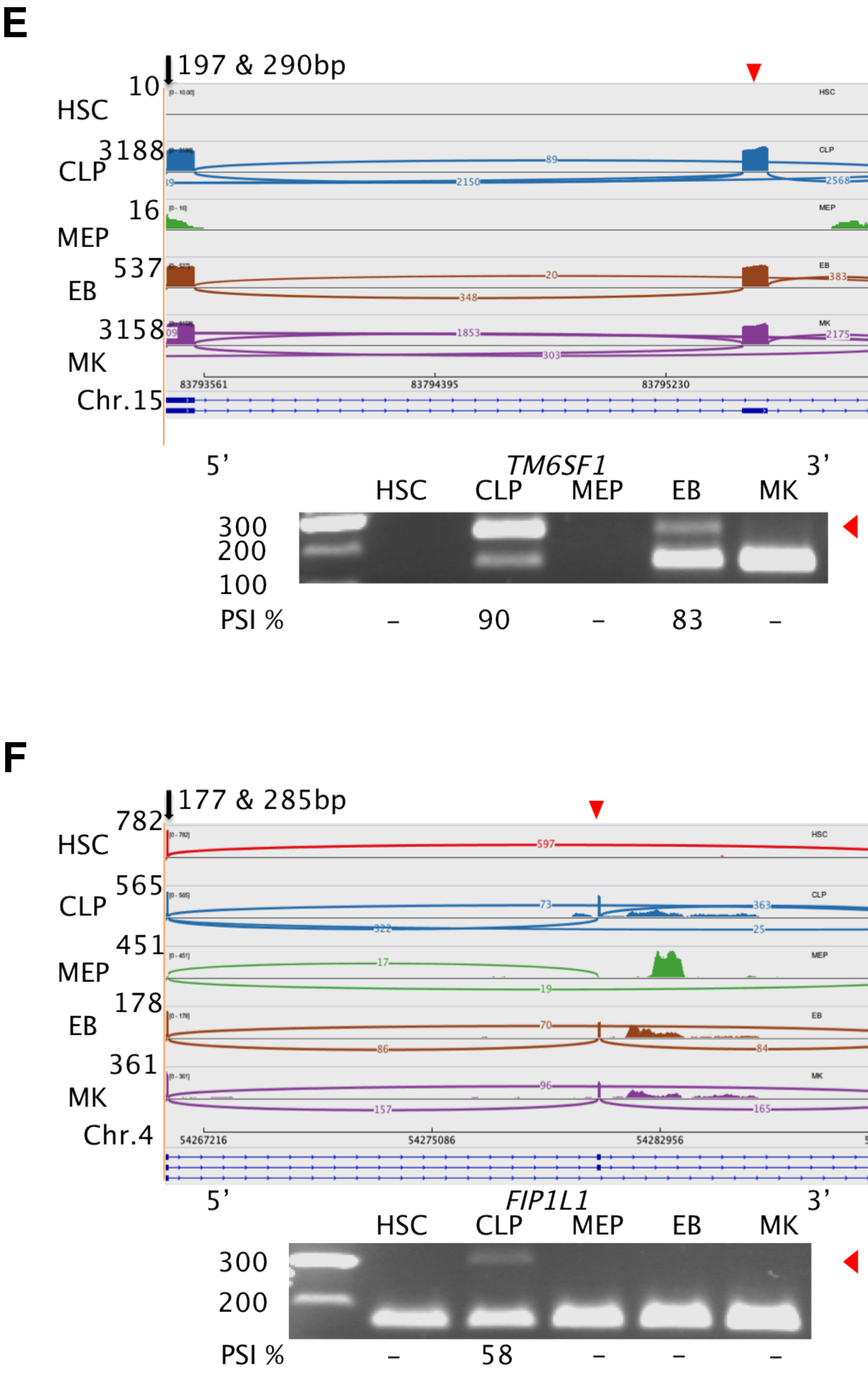


Figure S22: Distribution of putative single amino-acid insertion (left) or double amino-acid change (right) caused by the in-frame NAGNAG sequence, in alternative splice junctions with DSU (* represents stop codon).


A
177 & 316 bp
85
HSC
193
CLP
41
MEP
548
EB
236
MK
Chr.1
173772650
173773898
173775147
17377639
5'
CENPL
3'
HSC
CLP
MEP
EB
MK
300
200
PSI %
0
96
0
14
7
B
148 & 228 bp
255
HSC
202
CLP
120
MEP
945
EB
334
MK
Chr.13
36858472
36862712
36866953
36871194
3'
CCDC169
5'
HSC
CLP
MEP
EB
MK
200
100
PSI %
53
100
66
71
40


C
247 & 299 & 378 bp
83
HSC
158
CLP
451
MEP
123
EB
111
MK
Chr.1
114512785
114513753
114514721
114515690
5'
HIPK1
3'
HSC
CLP
MEP
EB
MK
400
300
200
PSI %
–
60
10
74
–
D
266& 449 bp
10
HSC
1841
CLP
3999
MEP
13873
EB
3192
MK
Chr.9
19118388
19119205
19120023
19120841
3'
PLIN2
5'
HSC
CLP
MEP
EB
MK
500
300
200
PSI %
–
43
80
90
–


E
197 & 290bp
10
HSC
3188
CLP
89
2150
2568
16
MEP
537
EB
20
348
383
3158
MK
1853
303
2175
83793561
83794395
83795230
83796065
Chr.15
5'
TM6SF1
3'
HSC
CLP
MEP
EB
MK
300
200
100
PSI %
–
90
–
83
–
F
177 & 285bp
782
HSC
597
565
CLP
73
363
322
25
451
MEP
17
19
178
EB
70
86
84
361
MK
96
157
165
Chr.4
54267216
54275086
54282956
54290826
5'
FIP1L1
3'
HSC
CLP
MEP
EB
MK
300
200
PSI %
–
58
–
–
–

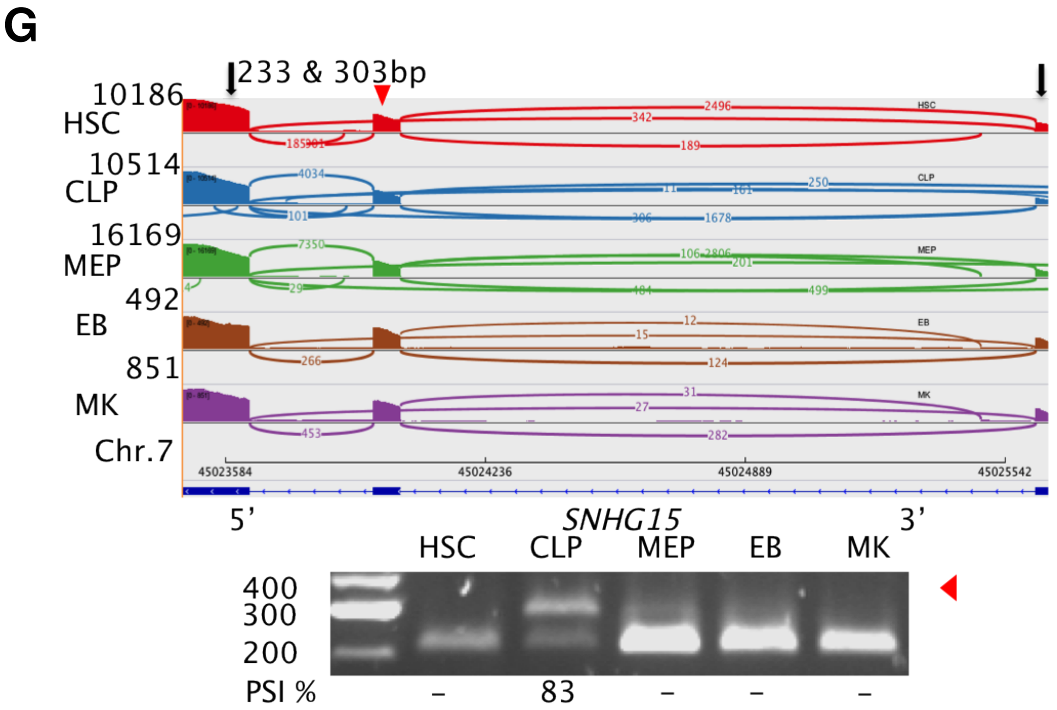


Figure S23: PCR validation of DSU events and RNA-seq PSI. For each splice junction the sashimi plot of RNA-seq reads in HSCs, CLPs, MEPs, EBs and MKs are shown in the top panel. The lines connecting two exons in the sashimi plot represent the splice junction with the number of reads supporting it within the line. The two black arrows on the top of the sashimi plot indicate the positions of primers with the estimated length of PCR product next to the left primer. PCR products visualized on gel are in the lower panel. Size markers are in base pairs. Red arrows indicate the PCR product that represent the DSU event and the position of the DSU in the sashimi plot. Values for RNA-seq PSI, for each cell type with DSU, were calculated using biological replicates and are displayed under the appropriate PCR fragment.

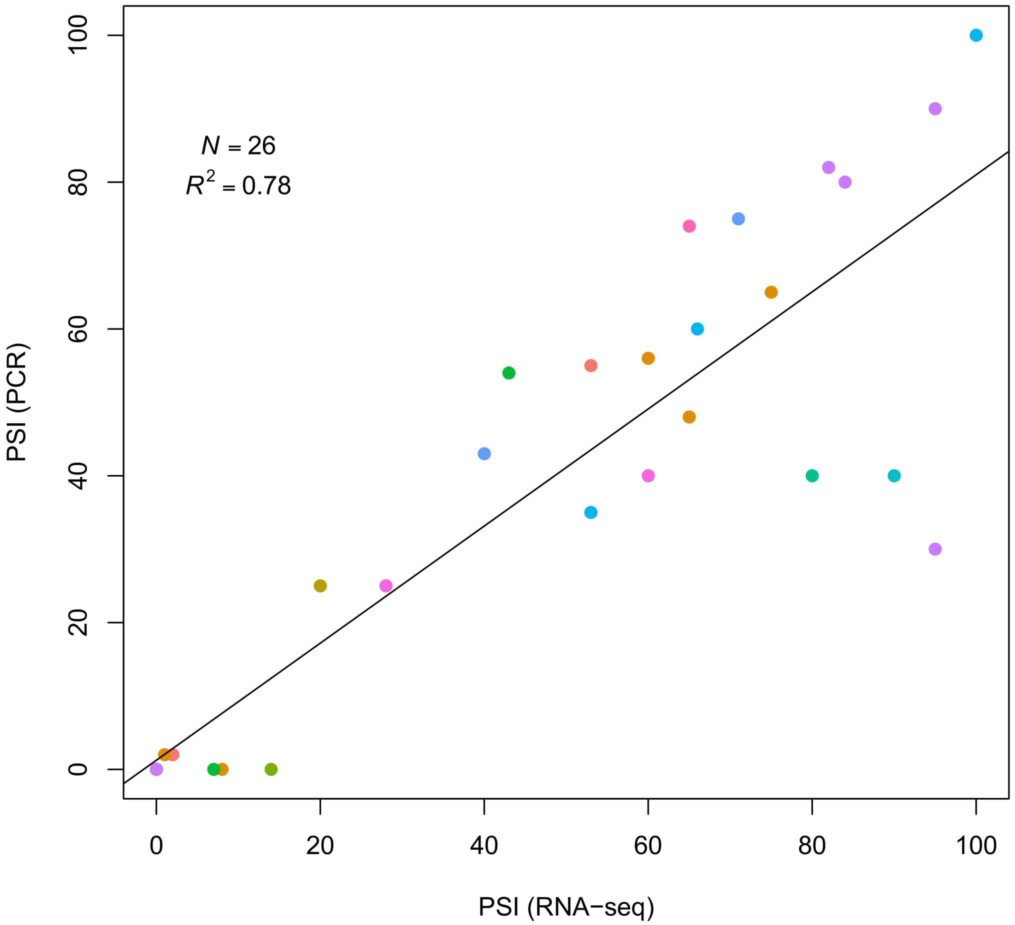


Figure S24: Correlation between PSI of the DSU from RNA-seq and the DSU PCRs (N=26, R^2=0.78). Each color represents a different PCR assay.

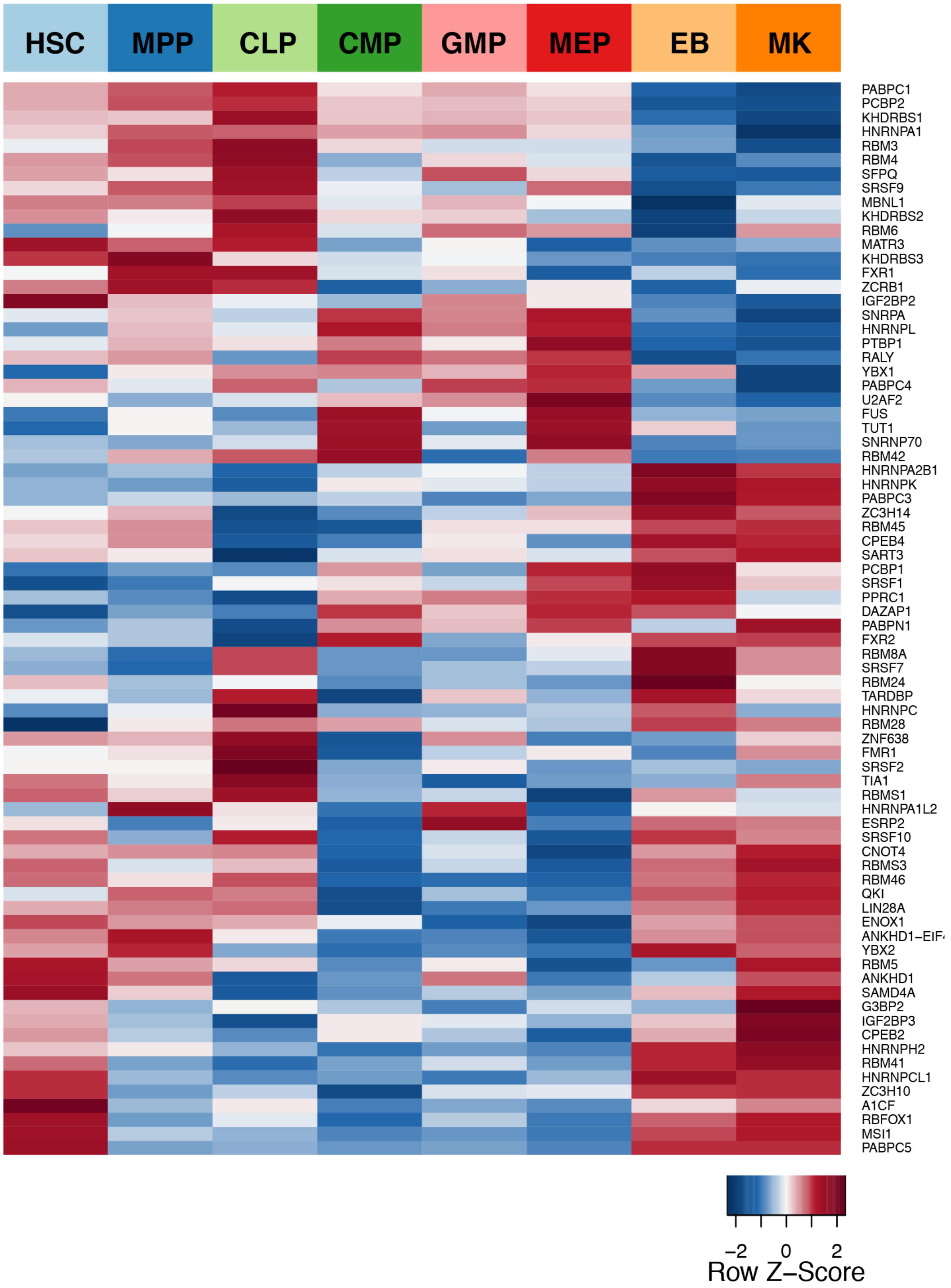


Figure S25: Normalized gene expression levels from MMSEQ analysis for RNA-binding proteins used in the RNA-binding motif enrichment analysis.

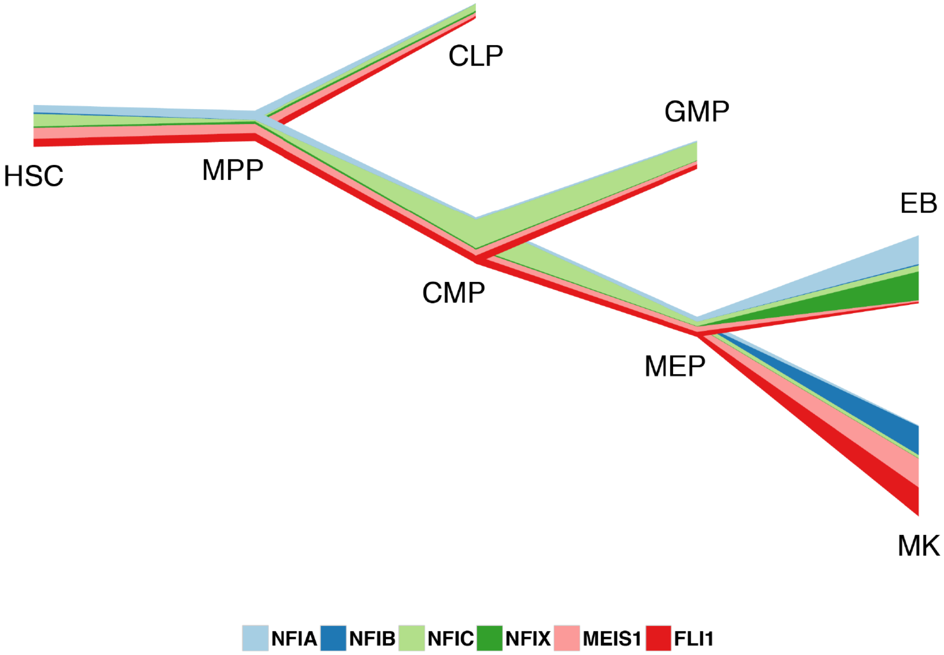


Figure S26: River plot illustrating the gene expression levels of NFI family members across all cell types. Width corresponds to expression level relative to the maximum expression level of the gene across all cell types. *MEIS1* and *FLI1* levels have been added for comparison with MK TFs. Note that MEIS1 and FLI1 bind the novel TSS of *NFIB* in MKs as shown in Fig. S27.

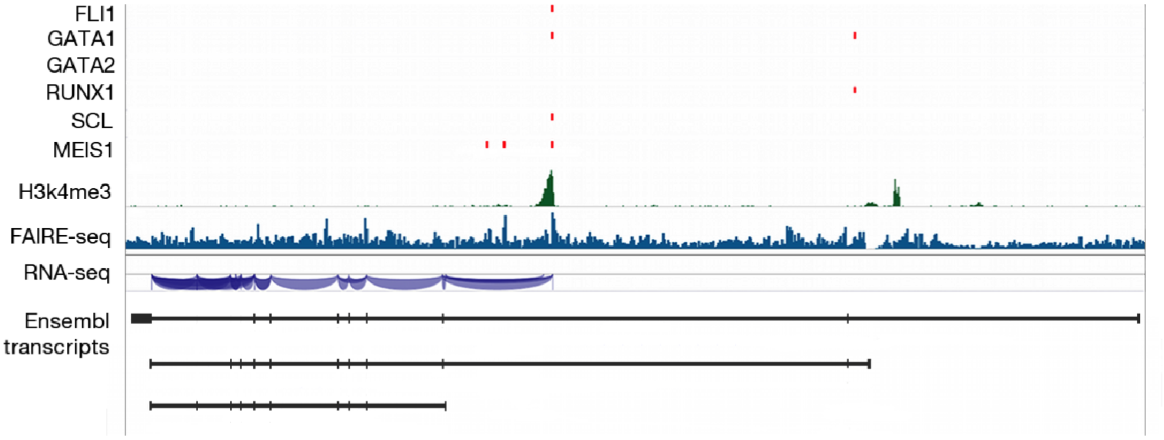


Figure S27: *NFIB* novel exon occupancy by FLI, GATA1, SCL and MEIS1 creates an open chromatin region marked by formaldehyde-assisted identification of regulatory elements (FAIRE) peak and a H3K4me3 peak.

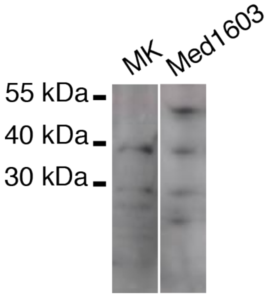


Figure S28: Expression of different protein isoforms of NFIB in MKs and the neuroblastoma MED1603 cell line. The longer isoform (*NFIB*-L) was present in the neuronal MED1603 cells, but absent from MKs.

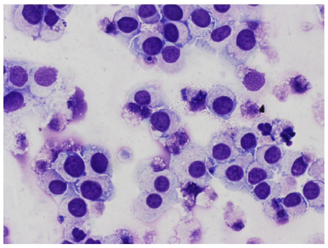
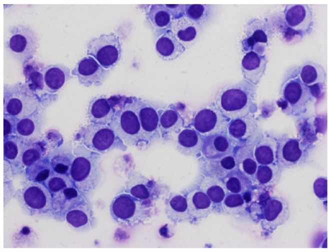
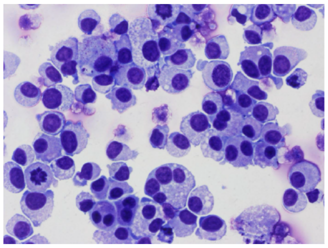
Image	**Morphology**
A shNFIB 	The majority of cells are intermediate sized immature cells with abundant vacuolated, granulated cytoplasm and eccentrically located nucleus with fine chromatin. Morphology of megakaryoblasts or promegakaryocytes.
B shNFIC 	Intermediate sized immature cells with abundant granulated cytoplasm and eccentrically located nucleus with fine chromatin. Morphology of megakaryoblasts or promegakaryocytes. Some more mature small mononuclear cells with condensed nuclear chromatin and greater cytoplasmic granulation: micromegakaryocytes.
C shCTRL 	Intermediate sized cells with abundant vacuolated, granulated cytoplasm and eccentrically located nucleus with fine-intermediate chromatin. Morphology of promegakaryocytes. Occasional binucleate and cells with lobated nuclei: morphology of more mature megakaryocytic cells.

Figure S29: Morphological analysis of MK cultures infected with shRNA against *NFIB* and *NFIC*.

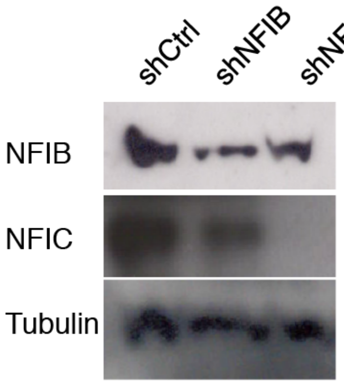


Figure S30: Down regulation of NFIB and NFIC in primary MKs. Western blot analysis of primary MKs transduced with shRNA pools against NFIB, NFIC and non-silencing control. Tubulin serves as loading control.

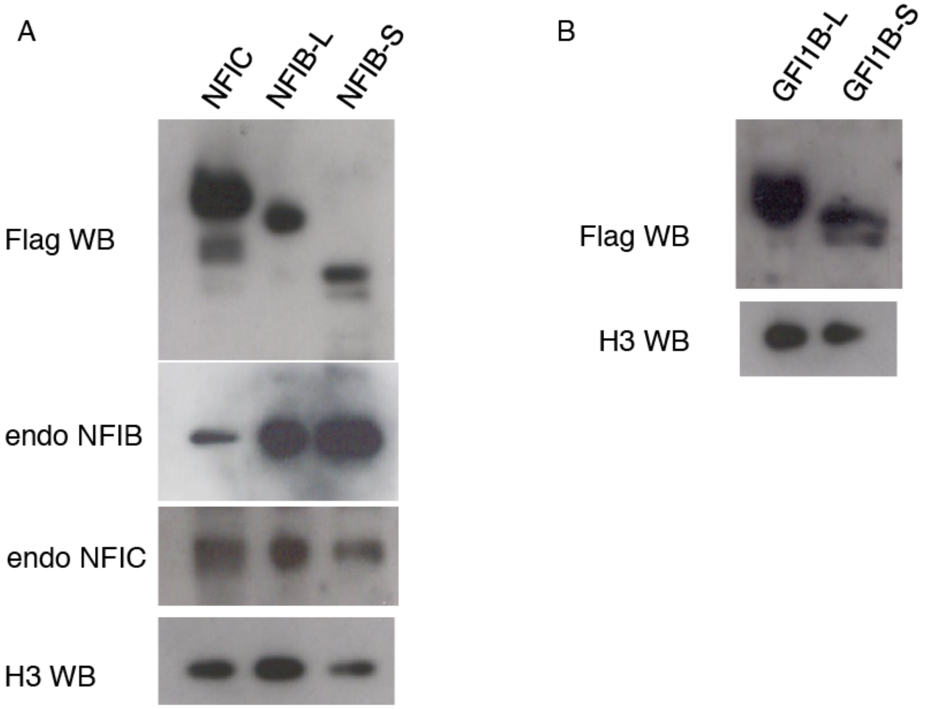


Figure S31: Transduction control of TAP tagged vectors for NFI (**A**) and GFI1B (**B**). K562 cells were transduced with the same MOI of lentiviral particles used in primary MKs and expanded. Nuclear extract were resolved by SDS Page and probed with M2 anti Flag antibody to reveal the TAP tagged proteins, endo NFIB and endo NFIC indicates the endogenous proteins. Histone 3 (H3) antibody was used as loading control.

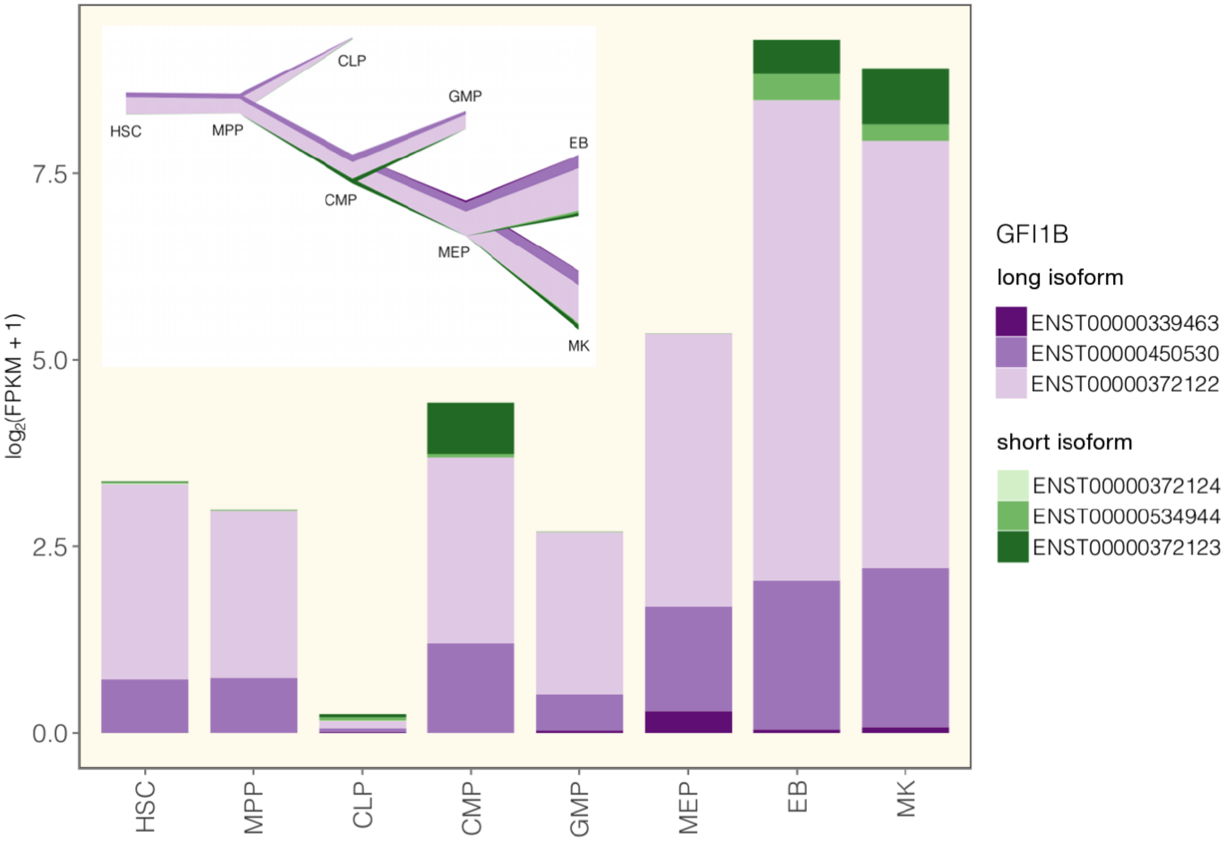


Figure S32: River plot and bar chart illustrating the expression of *GFI1B* isoforms across all samples. Transcript isoforms were grouped based on the length of the protein isoform they encode.

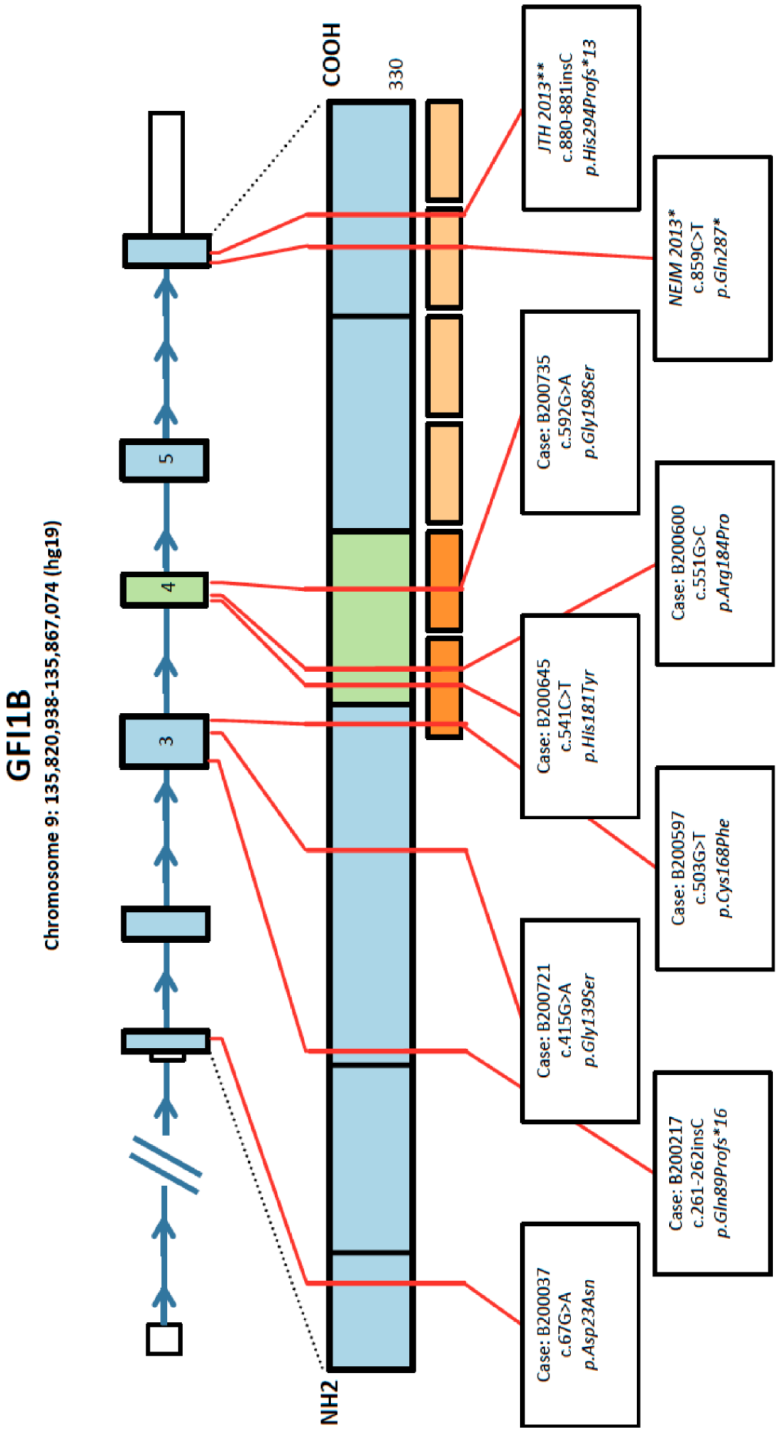
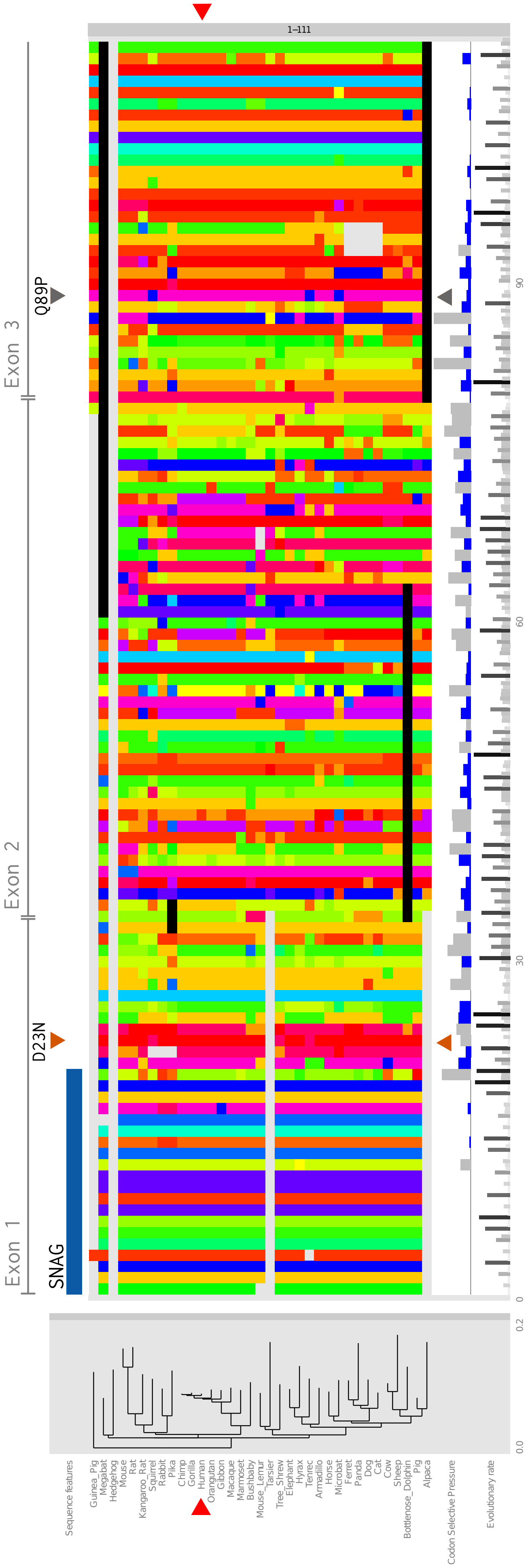


Figure S33: Summary of *GFI1B* structure and mutations previously reported in the literature and further identified in clinical BRIDGE cases. Orange boxes represent Zn-finger domains, with dark orange boxes representing Zn-fingers 1 and 2.


Exon 1
Exon 2
Exon 3
SNAG
D23N
Q89P
1–111
0
30
60
90
0.2
0.0
Sequence features
Guinea_Pig
Megabat
Hedgehog
Mouse
Rat
Kangaroo_Rat
Squirrel
Rabbit
Pika
Chimp
Gorilla
Human
Orangutan
Gibbon
Macaque
Marmoset
Bushbaby
Mouse_Lemur
Tarsier
Tree_Shrew
Elephant
Hyrax
Tenrec
Armadillo
Horse
Microbat
Ferret
Panda
Dog
Cat
Cow
Sheep
Bottlenose_Dolphin
Pig
Alpaca
Codon Selective Pressure
Evolutionary rate

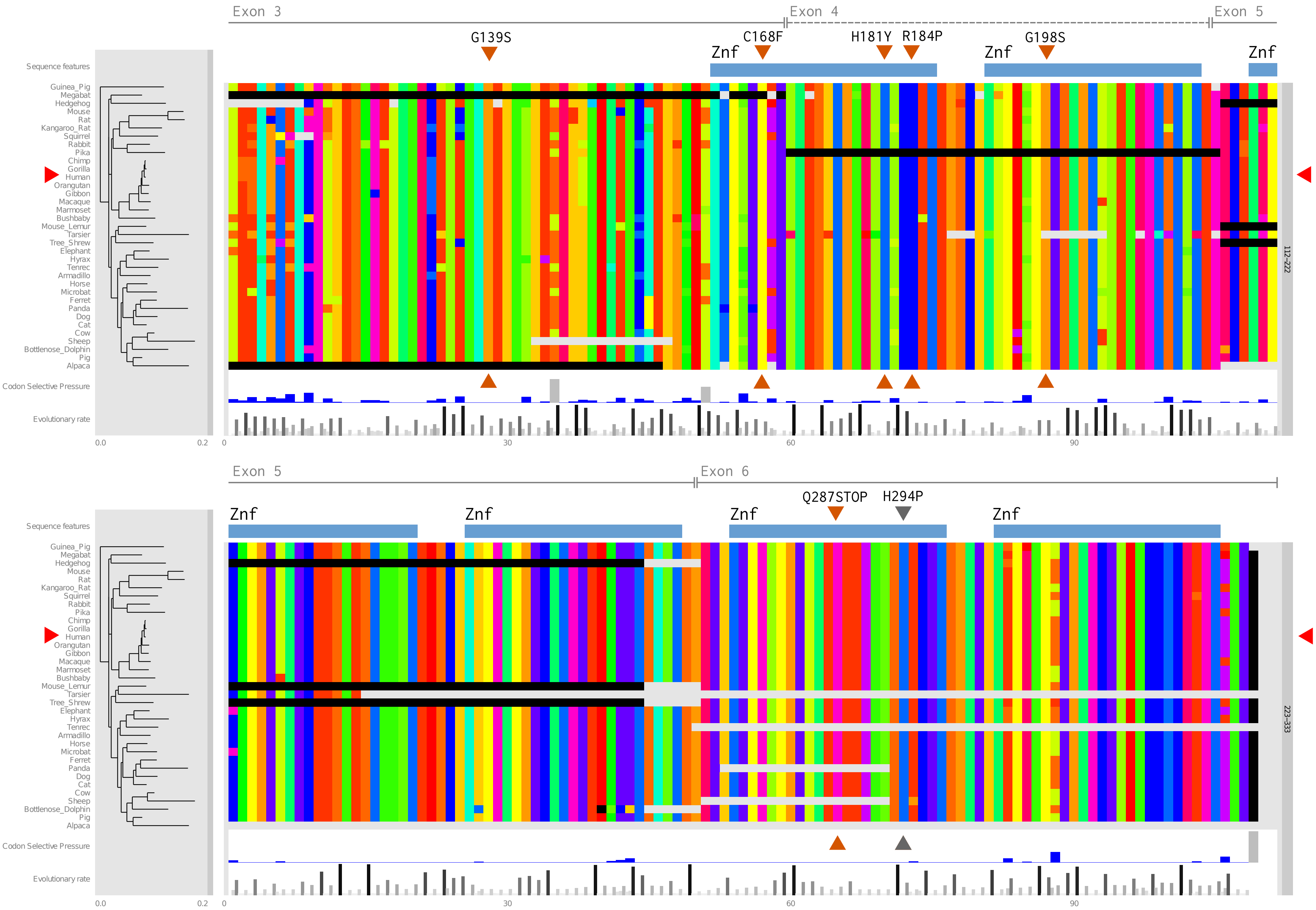

Exon 3
Exon 4
Exon 5
G139S
C168F
H181Y
R184P
G198S
Znf
Znf
Znf
Sequence features
Guinea_Pig
Megabat
Hedgehog
Mouse
Rat
Kangaroo_Rat
Squirrel
Rabbit
Pika
Chimp
Gorilla
Human
Orangutan
Gibbon
Macaque
Marmoset
Bushbaby
Mouse_Lemur
Tarsier
Tree_Shrew
Elephant
Hyrax
Tenrec
Armadillo
Horse
Microbat
Ferret
Panda
Dog
Cat
Cow
Sheep
Bottlenose_Dolphin
Pig
Alpaca
Codon Selective Pressure
Evolutionary rate
0.0
0.2
0
30
60
90
112–222
Exon 5
Exon 6
Q287STOP
H294P
Znf
Znf
Znf
Znf
Sequence features
Guinea_Pig
Megabat
Hedgehog
Mouse
Rat
Kangaroo_Rat
Squirrel
Rabbit
Pika
Chimp
Gorilla
Human
Orangutan
Gibbon
Macaque
Marmoset
Bushbaby
Mouse_Lemur
Tarsier
Tree_Shrew
Elephant
Hyrax
Tenrec
Armadillo
Horse
Microbat
Ferret
Panda
Dog
Cat
Cow
Sheep
Bottlenose_Dolphin
Pig
Alpaca
Codon Selective Pressure
Evolutionary rate
0.0
0.2
0
30
60
90
223–333

Figure S34: Evolutionary analysis of the coding region of the canonical transcript of *GFI1B*. The alignment of human and 34 other eutherian orthologs is shown, coloured by amino acids and with amino acid coordinates. To the left is the evolutionary tree relating the sequences, with human highlighted (red pointer). Sequence features were mapped using the human sequence and are shown above the alignment, including the locations of 7 missense mutations (orange pointers) and 2 frameshift mutations (grey). Site-wise estimates of selective pressure and evolutionary rates are given below the alignment. Selective pressure is shown using the relative non-synonymous:synonymous rate ratio dN/dS: values near 0 indicate strong purifying selection, with blue coloring used to indicate statistical significance. Grey bars indicate selection not significantly different from neutrality (dN/dS=1); no sites under positive selection were detected. Overall evolutionary rate is shown using the site-wise relative rate of DNA substitution, with bars shaded more darkly with increasing rate.

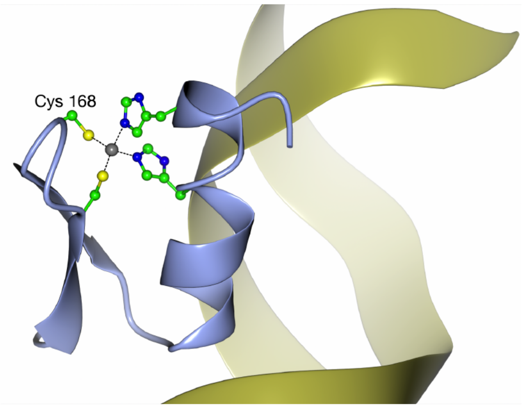


Figure S35: Three-dimensional model of the first Zn-finger of GFI1B with the cysteine (Cys) at residue 168 as green-yellow ball and stick. Cysteine and histidine residues coordinating the Zn ion (light blue) are shown. The Zn-finger domain is illustrated in complex with the major groove of DNA in yellow.

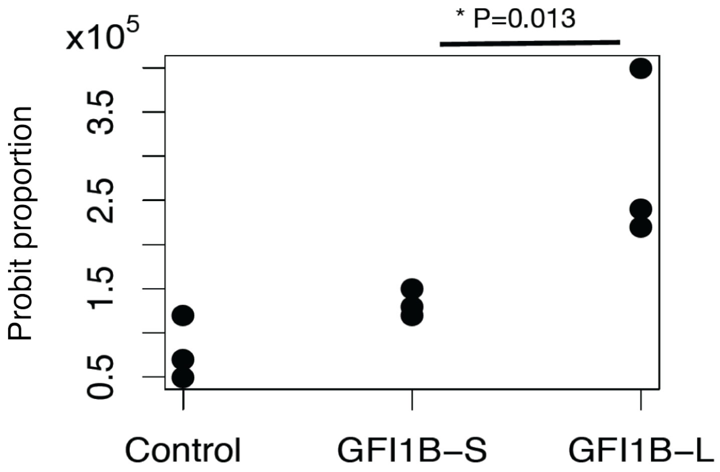


Figure S36: Effect of GFI1B isoforms on cell proliferation. Number of cells in MK cultures at day 10 after infection with lentiviral vectors.

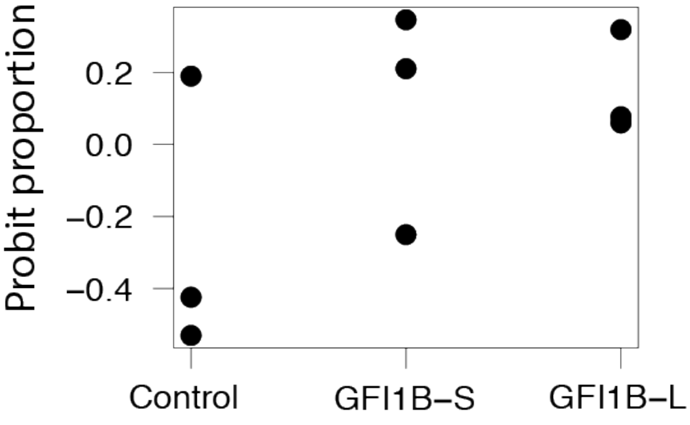


Figure S37: No differences in differentiation were observed with overexpression of the long (L) and short (S) *GFI1B* isoforms by over-expression within MKs. Proportion of CD41a (ITGA2B) and CD42b (GPIBA) double positive MKs in culture at day 10 after infection with lentiviral vectors expressing CTRL, GFI1B-L and GFI1B-S (probit proportion of the double positive MKs after adjusting for batch effects).

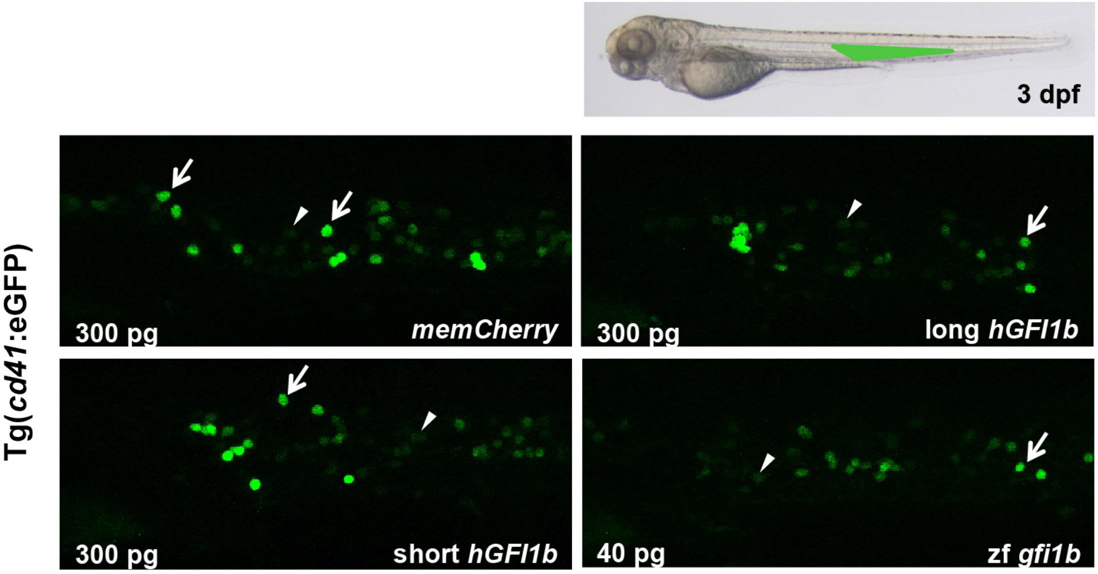


Figure S38: Over-expression of human *GFI1B* and zebrafish gfi1b in zebrafish embryos. Transgenic zebrafish *Tg(cd41:EGFP)* embryos were injected with long or short version of human GFI1B mRNA, zebrafish gfi1b mRNA (zf)-gfi1b or control membrane Cherry (memCherry) mRNA. At 3 days post fertilization, the region of caudal hematopoietic tissue (CHT, green) was viewed under the confocal microscope. The number of thrombocytes (arrows) and hematopoietic stem cells (arrowheads) was assessed. Representative images of CHT region are shown. All embryos are oriented with anterior to the left and dorsal to the top.

Supplementary tables

S1: Alignment statistics

S2: Primers and probe sequences for quantitative PCR BioMark assays

S3: Enriched transcript biotypes in each polytomous class

S4: Cell-type specific genes

S5: Cell-type specific transcripts

S6: 20 most enriched GO terms within cell-type specific genes

S7: 20 most enriched GO terms within cell-type specific transcripts

S8: GO and pathway enrichment analysis for genes with novel splice junctions

S9: PCR primers information for the validation of novel splice junctions and DSU

S10: Concordance of validation of novel splice junctions

S11: Number of genes with DSU and alternative splicing sets

S12: GO and pathway enrichment analysis for genes with DSU

S13: RNA-binding protein motifs and cluster information

S14: BRIDGE Bleeding and Platelet Disorders consortium members

S15: Genotypes and clinical phenotypes of BRIDGE cases with GFI1B variants