Full-length, single-cell RNA-sequencing of human bone marrow subpopulations reveals hidden complexity

Marcel O. Schmidt[1], Anne Deslattes Mays[2], Megan E. Barefoot[1], Anna T. Riegel[1], Anton Wellstein[1]

[1]Lombardi Comprehensive Cancer Center and Department of Oncology, Georgetown University, Washington, DC, [2]Science and Technology Consulting, LLC: Farmington, CT

Abstract

Bone marrow progenitor cell differentiation has frequently been used as a model for studying cellular plasticity and cell-fate decisions. Recent analysis at the level of single-cells has expanded knowledge of the transcriptional landscape of human hematopoietic cell lineages. Using single-molecule real-time (SMRT) full-length RNA sequencing, we have previously shown that human bone marrow lineage-negative (Lin-neg) cell populations contain a surprisingly diverse set of mRNA isoforms. Here, we report from single cell, full-length RNA sequencing that this diversity is also reflected at the single-cell level. From fresh human bone marrow unselected and lineage-negative progenitor cells were isolated by droplet-based single-cell selection (10xGenomics). The single cell-derived mRNAs were analyzed by full-length SMRT and short-read sequencing. In both samples we detected an average of 8000 different genes using short-read sequencing. Differential expression analysis arranged the single-cells of the total bone marrow into only four clusters whereas the Lin-neg population was much more diverse with nine clusters. mRNA isoform analysis of the single-cell populations using full-length sequencing revealed that Lin-neg cells contain on average 24% more novel splice variants than the total bone marrow cells. Interestingly, among the most frequent genes expressing novel isoforms were members of the spliceosome, e.g. HNRNPs, DEAD box helicases and SRSFs. Mapping the isoforms from all genes to the cell type clusters revealed that total bone marrow cells express novel isoforms only in a small subset of clusters. On the other hand, lineage-negative progenitor cells expressing novel isoforms

were present in nearly all subpopulations. In conclusion, on a single-cell level lineage-negative cells express a higher diversity of genes and more alternatively spliced novel isoforms suggesting that cells in this subpopulation are poised for different fates.

Introduction

In 2001, the International Human Genome Sequencing Consortium published that a typical human gene contains nine exons (Consortium 2004). During posttranscriptional editing of the nascent transcripts the introns are removed and exons spliced together to generate the coding or messenger RNA (mRNA). Alternative splicing of exons increases the diversity of mRNA isoforms, facilitating the generation of a multitude of protein variants with divergent functions from a single gene locus. Well-known examples for alternatively spliced gene products with distinct functions include isoforms of the *FAS* and *BCL-X* genes with either pro- or anti-apoptotic activities (Cheng et al. 1994; Boise et al. 1993). Also, the pro-angiogenic *VEGF* gene product can be alternatively spliced to produce an isoform with anti-angiogenic function (Nowak et al. 2008). In addition, multiple transcript isoforms are generated from the *TP53* tumor suppressor gene that are differently expressed between cancer and normal tissues (Vieler and Sanyal 2018).
Over the last decade technological advances have enabled gene expression analysis at the single-cell level. This has led to expanded knowledge of cellular diversity in terms of gene-expression. However, only few studies have aimed to characterize this diversity of splice variants at the single cell level (reviewed in Hardwick et al. 2019; Arzalluz-Luque and Conesa 2018). Several studies have evaluated isoform expression from single cells in mouse brain (Karlsson et al. 2017; Gupta et al. 2018) and in mouse B cells (Byrne et al. 2017). Nonetheless, similar studies have yet to be conducted with single-cells from human bone marrow, despite it being a continuously differentiating tissue that contains distinct subpopulations, and has a high turnover rate. The single-cell gene expression levels of human bone marrow-derived hematopoietic stem/progenitor cells have been described by several laboratories (Pellin et al. 2019; Velten et al. 2017). We have found that lineage-negative progenitor cells isolated from bulk bone marrow express a high

diversity of transcript isoforms including many novel splice variants not previously known (Deslattes Mays et al. 2019). Here, we investigate the extent to which individual cells within the total bone marrow (tot-BM) and the lineage negative (Lin-neg) populations exhibit this isoform diversity and whether distinct subpopulations in tot-BM and Lin-neg cells intersect based on transcript isoform usage. Using a droplet-based approach, we separately isolate tot-BM cells and Lin-neg cells, and sequence unfragmented libraries of full-length mRNA using Single Molecule Real-Time (SMRT) RNAseq technology on the PacBio platform.

Results

We extracted healthy donor human bone marrow tissues from discarded harvesting filters. From this total bone marrow cell preparation (tot-BM) we enriched for lineage-negative progenitor cells (Lin-neg) by magnetic selection as described earlier (Deslattes Mays et al. 2019). We then analyzed total and Lin-neg cell populations by droplet-based single cell RNA sequencing (10xGenomics). In order to increase the number of mRNA molecules detectable per cell we reduced the cell input to approximately 500 cells for each experiment. After single-cell selection, the barcoded cDNA libraries were equally divided and each pool was analyzed in parallel by short-read sequencing (Illumina) and by single-molecule real-time (SMRT) full-length RNA sequencing (Pacific Biosciences). The short-read sequencing results were used to cluster the single cells based on gene expression and identify the different cell types present in each sample. Full-length mRNA sequencing results were used to identify the transcript isoforms present within individual cells in different clusters.

(Graphical abstract)

Short-read sequencing reveals greater diversity in lineage-negative subpopulation compared to total bone marrow populations

The 10X-Genomics' microfluidic device isolates each cell in a water/oil emulsion resulting in cell-based and barcode tagged libraries with each cDNA molecule containing a unique molecular identifier (UMI). We split the cDNA libraries into two equal pools, one of which was analyzed by short-read RNAseq of fragmented cDNAs (Illumina) and the other by SMRT full-length RNAseq (PacBio).

The short-read data was evaluated with the Seurat package and we detected 415 total BM (tot-BM) and 492 lineage-negative (Lin-neg) cells. While we found comparable gene numbers in each sample (7212 genes in tot-BM, 8758 in Lin-neg) the number of distinct mRNA molecules detected using the UMI varied by over two-fold, i.e. 57,322 for tot-BM and 133,744 for Lin-neg (**Table 1**). We clustered the cells based on their gene expression levels and identified five clusters in tot-BM and nine in Lin-neg cell populations. We identified the most differentially expressed genes in each cluster and annotated each cluster by cell type (**Figure 1, Tables 2).** Supplemental **Tables S1 and S2** show the top marker genes expressed in each cluster. Interestingly, in addition to the five clusters identified from tot-BM cells we found four more clusters in the Lin-neg cells. Additional clusters were CD34+ progenitor (cluster 3), early B-cells (clusters 6,8), and immature granulocytes and neutrophils (cluster 2).

To identify potential overlap of clusters, we combined both samples in-silico and re-clustered the samples. The re-clustered cells separated into 10 clusters and we annotated the cell types in each cluster (**Figure 2A, Table 2**). Supplemental **Table S3** and **Figure S1** show the top marker genes defining these cell type clusters. The combination analysis revealed one additional cluster 3 with more mature CD34, CD14 positive myeloid cells, which was populated by a majority of tot-BM cells**.** Also, cluster 5 with early megakaryocytes contained more tot-BM cells (**Figure 2**). In comparison, Lin-neg cells were the majority contributors to two distinct populations of early B-cells identified from the combination clustering analysis, expressing either immunoglobulin genes (cluster 7) or MHC II (HLA) genes (cluster 9). The majority of clusters contain more Lin-neg than tot-BM cells reflecting the diversity of Lin-neg cells (**Figure 2B, Table 3**).

Full-length RNA sequencing reveals higher rate of novel isoforms in Lin-neg cell populations

Using the parallel pool of barcoded and UMI labeled single-cell cDNA libraries, we performed full-length SMRT sequencing that avoids fragmentation of the cDNA. Using the scalable de novo isoform discovery workflow, Isoseq3, we detected approximately 40 million circular consensus reads. We required three read passes to create a high-fidelity consensus sequence. Furthermore, we filtered the consensus reads to contain both 3' and 5' adapters, cell barcodes, UMI sequences and poly A tails to ensure that the sequences represent full-length molecules. The de-novo assembly generated 260,000 and 340,000 transcripts for the tot-BM and Lin-neg samples, respectively (**Table 4)**. The full-length reads were then mapped to the reference transcriptome using the Isoseq3 and cDNA cupcake packages. Transcript isoforms were annotated using the SQANTI3 package filtering out intra-priming instances. We found transcripts from 7,670 and 9,720 genes and 14,781 and 23,101 unique transcript isoforms in tot-BM and Lin-neg, respectively (**Table 4**). The transcript isoforms were classified into four categories with the SQANTI3 package (Tardaguila et al. 2018): The "Full Splice Match" (FSM) indicates an exact match with the number of exons and splice junctions of the reference transcriptome; isoforms in the "Incomplete Splice Match" (ISM) category lack 5' and/or 3' exons; "Novel In Catalog" (NIC) isoforms use new combinations of known junctions and "Novel Not In Catalog" (NNIC) are novel isoforms with at least one new splice junction (**Figure 3A**).

Remarkedly, in the Lin-neg cells, we observed a significant 27% increase (from 8.7% (tot-BM) to 11% (Lin-neg)) in splice variants with novel combinations of known exons (NIC) as well as a 21% increase (from 8.35% to 10.07%) isoforms with at least one novel splice site (NNIC) (**Figure 3A**). Further subtyping of each isoform is shown in **Figure 3B** and **Table 5**. FSM is further subcategorized into single- (SE) and multi-exonic (ME) genes. ISM is subclassified as internal (IF), as 5' or 3' fragments (5'F, 3'F). NIC contains combinations of known splice sites (KS) or junctions (KJ). In addition, ISM, NIC and NNC categories are subcategorized as having an intron retention (IR) or not. Overall, Lin-neg cells are 29% more enriched for novel isoforms (both NIC and NNIC) with both known and novel junctions and splice sites. This includes enrichment of splice variants containing intron retentions (IR, from 5.22% to 6.75%), which can lead to nonsense-mediated decay (NMD) (Jacob and Smith 2017).

Genes involved in mRNA processing show a high frequency of novel isoforms

Observing an overall increase of novel isoforms in the Lin-neg population led us to identify individual genes expressing novel variants. For that, we filtered the genes for expression levels (≥20 full-length molecules) and sorted them by the highest ratio for novel isoforms for tot-BM and Lin-neg cells (**Table 6**). Surprisingly, we found among both populations novel isoforms of genes involved in RNA splicing, such as members of the Heterogeneous Nuclear Ribonucleoproteins (HNRNPs), DEAD-Box Helicase 5 (DDX5, **Figure 3 C, D)**, Serine And Arginine Rich Splicing Factor 5 (SRSF5, **Figure 3 C, D**), Splicing Factor 1 (SF1), Ribonuclease T2 (RNASET2), and a Pre-mRNA Processing Factor (PRPF40A). While both samples revealed novel HNRNP isoforms, novel splice variants of DDX5, PRPF40A, RNASET2 and SRSF5 were enriched in the Lin-neg sample. On the other hand, novel splice isoforms of SF1 predominated in tot-BM.

Among other genes expressing novel splice variants, we found the Ornithine Decarboxylase Antizyme 1 (OAZ1, involved in polyamine metabolism), LIM Domain Only 2 (LMO2) maintains pre-erythrocytes in immature state) and a histone-encoding gene (H2AFY). Only few genes expressed more novel isoforms in the tot-BM than in the Lin-neg sample. Among them were Profilin 1 (PFN1) that is involved in cytoskeleton maintenance (**Figure 3 C, D**), a member of the Major histocompatibility class II complex (HLA-DQB1) and an isocitrate dehydrogenase (IDH3G) involved in the citrate cycle.

From our cluster analysis, we found several differentially expressed alternative splicing-associated genes in Lin-neg versus tot-BM cells (**Figure 4**). Splicing factors of the DDX helicase(Bourgeois et al. 2016) (e.g. DDX5) and the SRSF family (Twyffels et al. 2011) (e.g. SRSF5 and SRSF7) are higher expressed by Lin-neg cells in clusters 1, 2, and 5 to 9 while downregulated in cluster 4. Similarly, the PI3K related kinase (SMG1) (Chen et al. 2017), is also upregulated in clusters 1,2, and 5 to 8 while downregulated in cluster 4. Another gene involved in mRNA processing such as HNRNPA1 (Chen et al. 2010) had a similar expression pattern (**Figure 4**). Of particular interest, SMG1 is involved in nonsense-mediated mRNA decay (NMD, (Powers et al. 2020)) and we have observed an overall increase of intron retention in Lin-neg cells **(Figure 3B).** This observation led us to investigate the expression of “poison exons” of SRSF genes, defined as ultra-conserved regions containing noncoding exons, that, cause NMD when spliced-in (Lareau et al. 2007b; Leclair et al. 2020; García-Moreno and Romão 2020; Jacob and Smith 2017)**.** Indeed, we found such sequences for SRSF2, 5, 6 and 7 when we entered the coordinates published in (Leclair et al. 2020). Expression of the poison exons were increased in the Lin-neg population. Lin-neg cells express more poison exons of SRSF1,3,5,6, and 7 than tot-BM cells. Only poison exons of SRSF2 and 4 are expressed equally between the populations (**Figure 5**).

Lin-neg cells with novel isoforms map to most cell type clusters

After identification of alternative splicing of isoforms by full-length RNA sequencing we then used the single-cell barcodes to assign them to cell type clusters classified by short-read RNA sequencing. This classification reduced the total transcript numbers to 34,000 and 54,000 full-length transcripts for tot-BM and Lin-neg, respectively. (**Table 7**). We then filtered for cells with at least five isoforms. Those isoforms were expressed in 144 tot-BM and 292 Lin-neg cells. We mapped these cells to the combined cell type clusters and found that tot-BM cells with full-length isoforms spread across a subset of clusters (**Figure 6A, left panel)**. Clusters with CD33+ myeloid progenitor cells (cluster 2), CD34+/CD14 myeloid cells (cluster 3), and immature granulocytes/neutrophils (cluster 4) were the clusters containing the most selected cells. Only 4 full-length transcripts of tot-BM cells were detected in cluster 0 containing early myeloid cells. In contrast, Lin-neg cells were more evenly distributed among the clusters. Only clusters 0 and 4 were somewhat underrepresented (**Figure 6A, right panel**). We then sorted the cells by isoform classification (FSM, ISM, NIC, and NNIC) and mapped the top 25 cells in each category to the cell type clusters (**Figure 6B)**. We found that Lin-neg cells containing all isoform categories were widely distributed in nearly all clusters. Clusters with dendritic (cluster 8), IGH-pos early B-cells (cluster 9) and early erythrocytes (cluster 5) did not contain Lin-neg cells with a majority of full splice matches. In contrast, novel isoforms (ISM, NIC, and

NNIC) were found in all clusters except in cluster 0 with early myeloid cells and cluster 4 with immature granulocytes/neutrophils (**Figure 6B, bottom panel**). This distribution pattern was in stark contrast to the tot-BM cells, which we found to be enriched only in a smaller subset of clusters. Tot-BM cells with a majority of reference matching isoforms (FSM and ISM) were observed in clusters 1, 2, 3, 5, 6, and 8, while cells with the highest content of novel isoforms were mostly in clusters 2, 3, and 4 (**Figure 6B, top panel)**. We then compared the isoform expressions of the tot-BM and Lin-neg samples in all cells for each cluster directly (**Figure 6C)**. In a subset of clusters, Lin-neg cells with novel isoforms were enriched relative to tot-BM cells. We found that cells in clusters 5, 7, 8, and 9 contained novel isoforms from the Lin-neg population (NIC, or NNIC) whereas tot-BM cells with novel isoforms were not significantly increased. Interestingly, in cluster 0 (early Myeloid cells) with a total of 213 cells, only 89 and 4 full-length isoforms of Lin-neg and tot-BM populations, respectively, were detected. (**Table 3, 7**). Overall, we conclude that Lin-neg cells more frequently express a greater diversity of novel isoforms in a subset of gene expression clusters.

mRNA processing genes have the highest frequency of novel isoforms and are expressed across multiple cell type clusters

Next, we searched for cluster-specific splice variants of single genes. DDX5, the helicase involved in RNA splicing, expressed in all clusters (see **Figure 4)** is highly alternatively spliced in Lin-neg cells (see **Figure 3C, D)** and we found full-length transcripts in clusters 1,2,3,6, and 7. Interestingly tot-BM cells express canonical isoforms whilst Lin-neg cell isoforms were mostly novel (**Figure 7A)**. Isoforms of SRSF5, which is also expressed in most clusters and differentially spliced in the Lin-neg population (see **Figure 4** and **3C, D)**, were detected in clusters 1, 3, and 5 (**Figure 7A)**. Notably, tot-BM expressed only very few SRSF5 molecules (1 to 3) half of which were novel transcript isoforms. In contrast, Lin-neg cells contained more transcripts (2 to 14), of which the majority were novel (NIC or NNIC) (**Figure 7A)**. Cells with transcripts isoforms of the Heterogeneous Nuclear Ribonucleoproteins (HNRNPM and HNRNPF) populated clusters 1, 3, 5, 6 and 1 to 5 respectively. While more isoforms were found in Lin-neg cells than tot-BM (28 vs

17 HNRNPM, and 34 vs 8 HNRNPF) the majority in both samples expressed NNIC isoforms (**Figure 7A, B)**. This is surprising and points to yet unknown novel isoforms in this gene family. PFN1, expressed at higher levels in tot-BM cells is also differentially spliced. Tot-BM cells in clusters 2, 3, 5, and 6 express mainly novel splice variants (NIC) whereas Lin-neg cells express only isoforms matching the reference (FSM) (**Figure 3C, D** and **7A**).

In summary, our data corroborates that on a single-cell level lineage-negative cells express not only a higher diversity of genes but also more alternatively spliced novel isoforms. From the single-cell analysis, we expand on this finding to reveal that tot-BM cells expressing novel isoforms were mostly located to the CD14+/CD34+ cell cluster 3. In contrast, Lin-neg cells expressing novel isoforms were present in nearly all subpopulations. Further, we discovered that many novel isoforms across all clusters were members of the spliceosome machinery, suggesting a previously unknown functional utility of the observed isoform heterogeneity in human bone marrow progenitor cells.

Discussion

Here we report a high-resolution transcriptional landscape of bone marrow populations both at the level of single-cells and full-length isoforms. We analyzed single-cells from total bone marrow and lineage-negative progenitor subpopulations isolated from the same donor with both short-read and full-length RNA sequencing technology. Previously, we have studied bulk mRNA from the same type of cell populations and found that lineage negative cells express more diverse splice variants including many novel isoforms than the total bone marrow (Deslattes Mays et al. 2019). We now report that this diversity is also reflected at single-cell resolution and not merely a by-product of greater cell type variation being captured in the bulk mRNA homogenate. Lin-neg cells analyzed by short-read sequencing express an expansive variety both in differential expression and in isoform diversity as shown by the number of clusters (9 in Lin-neg vs 5 in tot-BM) and mRNA molecules (133,700 Lin-neg vs 57,300 tot-BM) in a similar number of cells (492 Lin-neg vs 415 tot-BM cells). Full-length transcript analysis indicated an overall increase of novel isoforms. In addition, the Lin-neg cells contained clusters for a wide array of additional progenitor cell clusters such as progenitor cells for granulocytes, neutrophils, basophils, eosinophils, mast cells and B-lymphocytes as well as a subset of CD34-pos stem cells (**Table 2** and **Figure 1A**). This clustering reflects the greater diversity in Lin-neg cells. After merging both samples in silico and re-clustering, the tot-BM cells populated the additional clusters observed with Lin-neg cells (**Figure 2A**). Across the board, Lin-neg cells account for the majority composition in all clusters except for cluster 3 (more mature CD14+ myeloid cells) and cluster 5 (megakaryocytes), where tot-BM cells predominate, reflecting the more mature cell content. The higher diversity of Lin-neg cells indicates the higher plasticity of Lin-neg cells and negates the notion that this population is an amorphous pool of progenitor cells. Rather, our data suggest that the heterogenous Lin-neg progenitor cell subpopulations are poised for different fates.

Also, the single-cell full-length RNA sequencing analysis revealed that individual Lin-neg cells contain more novel isoforms than individual cells in the tot-BM. This extends our previous observation of RNA isoform differences between pools of these cell populations (Deslattes Mays et al. 2019). Interestingly, alternative splicing in Lin-neg cells is found in all subpopulations whereas in tot-BM it is more restricted and predominantly detected in myeloid cells. We were surprised to find spliceosome-associated genes among the genes with the most novel isoforms such as members of the DEAD-box helicases (DDXs), heterogeneous nuclear ribonucleoproteins (HNRNPs), splice regulatory (SR) proteins such as SRSF5, RNase T2 and Splicing Factor 1. In most of the Lin-neg cell clusters these spliceosomal genes including SMG1 were also upregulated relative to tot-BM. It is intriguing that genes involved in mRNA splicing were themselves alternatively spliced as we observed in many cell type clusters of the Lin-neg population. Interestingly, a recent publication (Lee et al. 2018) showed that after knockdowns of HNRNPA1 or DDX5, most alternative splicing events were in genes involved in RNA processing. Among them were SMG7, DDX, SRSF and HNRNP gene family members. Indeed, many recent reviews show that alternative splicing plays a vital role in the self-renewal, pluripotency and lineage specific differentiation of hematopoietic stem cells (Wong et al. 2018; Goldstein et al. 2017; Chen et al. 2014; Li et al. 2021).

Much research has been done on intron retention as a regulatory mechanism for gene expression during hematopoiesis. Intron retention may cause the emergence of a premature termination codon (PTC), which in turn can lead to nonsense-mediated decay (NMD) of mRNA. But NMD is not only an RNA surveillance mechanism in order to degrade mRNAs with nonsense mutations but also serves in an essential regulatory role in post-transcriptional gene expression control in vertebrates (Yi et al. 2020; García-Moreno and Romão 2020; Wong et al. 2018; Lykke-Andersen and Jensen 2015). Finally, previous research has shown that SR proteins are auto-regulated by alternative splicing coupled to nonsense-mediated decay (AS-NMD). SR protein genes comprise conserved regions containing noncoding exons, called ‘‘poison exons’’ (PEs), that introduce a PTC and target the mRNA for degradation. We found poison exons expressed predominantly in Lin-neg cells. This finding underscores the role of alternatively spliced poison exons as a potential regulator of gene expression during differentiation of hematopoietic stem cells as described by others (García-Moreno and Romão 2020; Lareau et al. 2007a; Leclair et al. 2020; Jacob and Smith 2017). Overall, our single-cell analysis of bone marrow populations revealed that lineage-negative cells express more diverse splice variants with higher content of novel isoforms. We observed novel isoforms in all Lin-neg cell type clusters and also found them in spliceosome-associated genes. Our findings suggest that isoform diversity by alternative splicing in Lin-neg cells is associated with distinct cell-fate decisions in each cell subpopulation. Future work is needed to explore how alternative splicing is regulating the differentiation of resident bone marrow cells. Especially, investigating the time-dependent expression of novel isoform during maturation of progenitor cells into terminally differentiated cells is of great interest. Furthermore, exploration of the functional activities associated with novel isoforms identified from this analysis could lead to increased understanding of hematopoietic cell differentiation.

Methods

Total and lineage negative bone marrow cells

The study was reviewed and considered as “exempt” by the Institutional Review Board of Georgetown University (IRB # 2002-022). All methods were carried out in accordance with relevant guidelines and regulations. Freshly harvested bone marrow tissue was collected from a single discarded healthy human bone marrow collection filter that had been de-identified. mononuclear cells were isolated by Ficoll gradient centrifugation. In order to select for lineage-negative cells, bone marrow mononuclear cells were negatively selected with an antibody cocktail containing antibodies against CD2, CD3, CD5, CD11b, CD11C, CD14, CD16, CD19, CD24, CD61, CD66b, and Glycophorin A (Stemcell Technologies, Vancouver, British Columbia, Canada).

Single cell cDNA library preparation

Initially 750 of each tot-BM and 750 Lin-neg cells were isolated and then added to the 10x Genomics Chromium controller according to the Chromium Single Cell 3′ Reagent Kits V3 User Guide for a final count of ca. 500 cells. We modified the amplification step (step 2), reducing the cycles from the recommended 14 to 10 in order to capture more rare transcripts. The amplified cDNA library was then spit. 25% was used for Illumina short-read library preparation and 75% for the PacBio library preparation.

Short-read RNA-seq

For the short-read library preparation the chromium user guide was continued and the final fragmented cDNA libraries were sent to Novagene (Sacramento, CA) and sequenced on an Illumina Hiseq 4000 with a depth of 100 million paired reads per sample and a read length of 150nt. The resulting raw reads (FASTQ) were then de-multiplexed using cellranger software (10xgenomics). The reads were filtered for highly diverse UMI in order to remove bulk non-single cell-derived transcripts. The reads were then clustered with the Cell-Loupe browser (10xGenomics) to ascertain the quality of samples. The resulting non-normalized data were then fed into the Seurat pipeline (Stuart et al. 2019) for more detailed analysis.

Full-length RNA-seq

The cDNA libraries containing sample index, UMI, and the cell barcode from step 2 of the 10xpipeline was sent to University of Maryland's Institute for Genome Sciences for full-length sequencing on the PacBio Sequel 2 platform. The two sample libraries were pooled with two other libraries and run for 30 hours. The resulting circular consensus (CSS) reads were demultiplexed, filtered for consensus within 3 passes and PacBio's Isoseq3 (https://github.com/PacificBiosciences/IsoSeq) pipeline was followed with modifications for single cell data (Cupcake by Elizabeth Tseng https://github.com/Magdoll/cDNA_Cupcake, SQANTI3: https://github.com/ConesaLab/SQANTI3). In short 3' and 5' primers were detected and removed (Lima), then UMIs and Cell barcodes were detected (clip_out_UMI_cellBC.py), finally the polyA tail was detected and artificial concatemers were removed (isoseq3 refine). Then the reads were aligned to the genome (minmap2) and unique transcripts were collapsed (collapse_isoforms_by_sam.py). The reads were then annotated with sqanti3_qc.py, artefacts were filtered out (sqanti3_RulesFilter.py), and the high-quality transcripts were combined with the UMI and cell barcodes for the final report (collate_FLNC_gene_info.py, UMI_BC_error_correct.py)

Single cell cluster analysis (Seurat**,** (Stuart et al. 2019)**)**

Single cell analysis followed vignettes posted in github: https://github.com/satijalab/seurat. Briefly, count matrix output from Cellranger (10xgenomics) was read into Seurat software on R. Quality control of the count matrix was performed by filtering out single cells with more than 25% mitochondrial mRNA content and less than 200 UMIs as those represent empty droplets. The UMI counts were normalized, log transformed and scaled by linear transformation to equalize mean and variance across genes. Principal component analysis and determination of dimensionality of the data was performed and the data was visualized by clustering the cells using non-linear dimensional reduction (tSNE or UMAP). In order to annotate the cell types of each cluster differentially expressed features were used as find biomarker. For differential analysis the Seurat objects for each sample (tot-BM and Lin-neg) were merged and the differential gene expression analysis was repeated. Barcodes from each cluster were identified and used to select and identify full-length reads for isoform analysis.

Code availability—custom code is available upon request.

Figure legends

Figure 1. Lineage-negative cells have greater diversity than total bone marrow cells. (**A**) Single-cell short-read expression analysis with the Seurat R package reveal 5 clusters for tot-BM cells (left panel) and 9 for Lin-neg cells (right panel). Clusters are annotated according to marker gene expression (Wu et al. 2016; Uhlén et al. 2015; Pellin et al. 2019; Velten et al. 2017). (**B**) Heatmaps of the top 10 regulated genes for each cluster. (**C**) Violin expression plots of the top cell type marker for each cluster. The full names of genes indicated as acronyms are listed in Tables S1 and S2

Figure 2. Most clusters of the combined single-cell expression analysis contain more Lin-neg than tot-BM cells. (**A**) tSNE plot of combined cluster analysis colored by sample (left panel) or by cluster (right panel). Clusters are annotated according to marker gene expression (Wu et al. 2016; Uhlén et al. 2015; Pellin et al. 2019; Velten et al. 2017). (**B**) Total number (left) and fraction of cells per cluster (right). Chi-square tests for total BM and Lin-neg cells number distribution in clusters show highly significant distributions ($p < 0.0001$)

Figure 3. Lin-neg cells express more novel isoforms than tot-BM cells. (**A**) Major isoform categories for each sample. Consensus isoforms are either “Full Splice Match” (FSM) or ”Incomplete Splice Match” (ISM). Novel isoforms are either “Novel In Catalog” (NIC) or “Novel Not In Catalog” (NNIC). (**B**) Subcategories of alternatively spliced isoforms. Multi and single exon are subcategories of FSM. 3’ fragment, 5’ fragment, and internal fragment are subcategories of ISM. The NIC category is subdivided into known junctions and known splice sites. NNIC contains novel splice sites. ISM, NIC and NNC categories may also contain an intron retention. (**C**) Isoform categories and relative content for DDX5, SRSF5 and PFN1. (**D**) UCSC genome browser image of unique isoforms detected in each sample. Isoforms are labeled with categories and subcategories. Multi exon (ME), single exon (SE), 3’ fragment (3’F), 5’ fragment (5’F), internal fragment (IF), known junctions (KJ), known splice sites (KS), intron retention (IR). Consensus translation starts and stops are indicated. Chi-square tests for total BM and Lin-neg isoform distribution show highly significant distributions ($p < 0.0001$) for A, B and C (DDX5, PFN1), $p=0.0054$ for SRSF5)

Figure 4. Spliceosome-associated genes are alternatively spliced and have higher expression in Lin-neg cells in most clusters Dot plots of expression levels of alternatively spliced genes detected by short- read sequencing arranged in clusters. Size of dots corresponds to percentage of cells expressing the gene, Color tint reflects average expression level.

Figure 5. Spliceosome-associated transcripts contain poison exons. Isoforms detected by full-length sequencing for SRSF1 to 7. Poison exons are highlighted in grey. Analysis from the UCSC browser

Figure 6. Lin-neg cells expressing novel isoforms are found in most clusters. (**A**) tSNE plot of tot-BM cells (left) and Lin-neg cells (right). Cells containing full-length reads filtered for at least 5 reads per gene are highlighted in blue (tot-BM) and in red (Lin-neg). (**B**) tSNE plots depict the top 25 cells of each isoform category of the tot-BM (highlighted in blue, top panels) and the Lin-neg population (highlighted in red, bottom panels). (**C**) Bar graph showing expression of isoform categories in Lin-neg cells by cell type cluster. Full Splice Match = FSM, Incomplete Splice Match = ISM, Novel In Catalog = NIC, Novel Not In Catalog = NNIC.

Figure 7 mRNA-processing genes in Lin-neg cells express novel isoforms in most cell type clusters. (**A**) bar graphs showing full-length reads (FL-reads) of isoform categories for DDX5, SRSF5, HNRNPM, HNRNPF, and PFN1. (**B**) USCS genome browser image of full-length isoforms of HNRNPF and HNRNPM. Each isoform is labeled with its respective splicing category. Full Splice Match = FSM, Incomplete Splice Match = ISM, Novel In Catalog = NIC, Novel Not In Catalog = NNIC.

References

Arzalluz-Luque Á, Conesa A. 2018. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biol* **19**: 110.

Boise LH, González-García M, Postema CE, Ding L, Lindsten T, Turka LA, Mao X, Nuñez G, Thompson CB. 1993. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* **74**: 597–608.

Bourgeois CF, Mortreux F, Auboeuf D. 2016. The multiple functions of RNA helicases as drivers and regulators of gene expression. *Nat Rev Mol Cell Bio* **17**: 426–438.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun* **8**: 16027.

Cheng J, Zhou T, Liu C, Shapiro J, Brauer M, Kiefer M, Barr P, Mountz J. 1994. Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science* **263**: 1759–1762.

Chen J, Crutchley J, Zhang D, Owzar K, Kastan MB. 2017. Identification of a DNA Damage–Induced Alternative Splicing Pathway That Regulates p53 and Cellular Senescence Markers. *Cancer Discov* **7**: 766–781.

Chen L, Kostadima M, Martens JHA, Canu G, Garcia SP, Turro E, Downes K, Macaulay IC, Bielczyk-Maczynska E, Coe S, et al. 2014. Transcriptional diversity during lineage commitment of human blood progenitors. *Science* **345**: 1251033.

Chen M, Zhang J, Manley JL. 2010. Turning on a Fuel Switch of Cancer: hnRNP Proteins Regulate Alternative Splicing of Pyruvate Kinase mRNA. *Cancer Res* **70**: 8977–8980.

Consortium IHGS. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

García-Moreno JF, Romão L. 2020. Perspective in Alternative Splicing Coupled to Nonsense-Mediated mRNA Decay. *Int J Mol Sci* **21**: 9424.

Goldstein O, Meyer K, Greenshpan Y, Bujanover N, Feigin M, Ner-Gaon H, Shay T, Gazit R. 2017. Mapping Whole-Transcriptome Splicing in Mouse Hematopoietic Stem Cells. *Stem Cell Rep* **8**: 163–176.

Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. 2018. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* **36**: 1197–1202.

Hardwick SA, Joglekar A, Flicek P, Frankish A, Tilgner HU. 2019. Getting the Entire Message: Progress in Isoform Sequencing. *Frontiers Genetics* **10**: 709.

Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**: 1043–1057.

Karlsson K, Lönnerberg P, Linnarsson S. 2017. Alternative TSSs are co-regulated in single cells in the mouse brain. *Mol Syst Biol* **13**: 930.

Lareau LF, Brooks AN, Soergel DAW, Meng Q, Brenner SE. 2007a. Alternative Splicing in the Postgenomic Era. *Adv Exp Med Biol* **623**: 190–211.

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007b. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446**: 926–929.

Leclair NK, Brugiolo M, Urbanski L, Lawson SC, Thakar K, Yurieva M, George J, Hinson JT, Cheng A, Graveley BR, et al. 2020. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. *Mol Cell* **80**: 648-665.e9.

Lee YJ, Wang Q, Rio DC. 2018. Coordinate regulation of alternative pre-mRNA splicing events by the human RNA chaperone proteins hnRNPA1 and DDX5. *Gene Dev* **32**: 1060–1074.

Li Y, Wang D, Wang H, Huang X, Wen Y, Wang B, Xu C, Gao J, Liu J, Tong J, et al. 2021. A splicing factor switch controls hematopoietic lineage specification of pluripotent stem cells. *Embo Rep* **22**: e50535.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Bio* **16**: 665–677.

Deslattes Mays A, Schmidt M, Graham G, Tseng E, Baybayan P, Sebra R, Sanda M, Mazarati J-B, Riegel A, Wellstein A. 2019. Single-Molecule Real-Time (SMRT) Full-Length RNA-Sequencing Reveals Novel and Distinct mRNA Isoforms in Human Bone Marrow Cell Subpopulations. *Genes-basel* **10**: 253.

Nowak DG, Woolard J, Amin EM, Konopatskaya O, Saleem MA, Churchill AJ, Ladomery MR, Harper SJ, Bates DO. 2008. Expression of pro- and anti-angiogenic isoforms of VEGF is differentially regulated by splicing and growth factors. *J Cell Sci* **121**: 3487–3495.

Pellin D, Loperfido M, Baricordi C, Wolock SL, Montepeloso A, Weinberg OK, Biffi A, Klein AM, Biasco L. 2019. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat Commun* **10**: 2395.

Powers KT, Szeto J-YA, Schaffitzel C. 2020. New insights into no-go, non-stop and nonsense-mediated mRNA decay complexes. *Curr Opin Struc Biol* **65**: 110–118.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive Integration of Single-Cell Data. *Cell* **177**: 1888-1902.e21.

Tardaguila M, Fuente L de la, Marti C, Pereira C, Pardo-Palacios FJ, Risco H del, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28**: 396–411.

Twyffels L, Gueydan C, Kruys V. 2011. Shuttling SR proteins: more than splicing factors. *Febs J* **278**: 3246–3255.

Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, et al. 2017. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* **19**: 271–281.

Vieler M, Sanyal S. 2018. p53 Isoforms and Their Implications in Cancer. *Cancers* **10**: 288.

Wong ACH, Rasko JEJ, Wong JJ-L. 2018. We skip to work: alternative splicing in normal and malignant myelopoiesis. *Leukemia* **32**: 1081–1093.

Yi Z, Sanjeev M, Singh G. 2020. The Branched Nature of the Nonsense-Mediated mRNA Decay Pathway. *Trends Genet* **37**: 143–159.

Graphical abstract

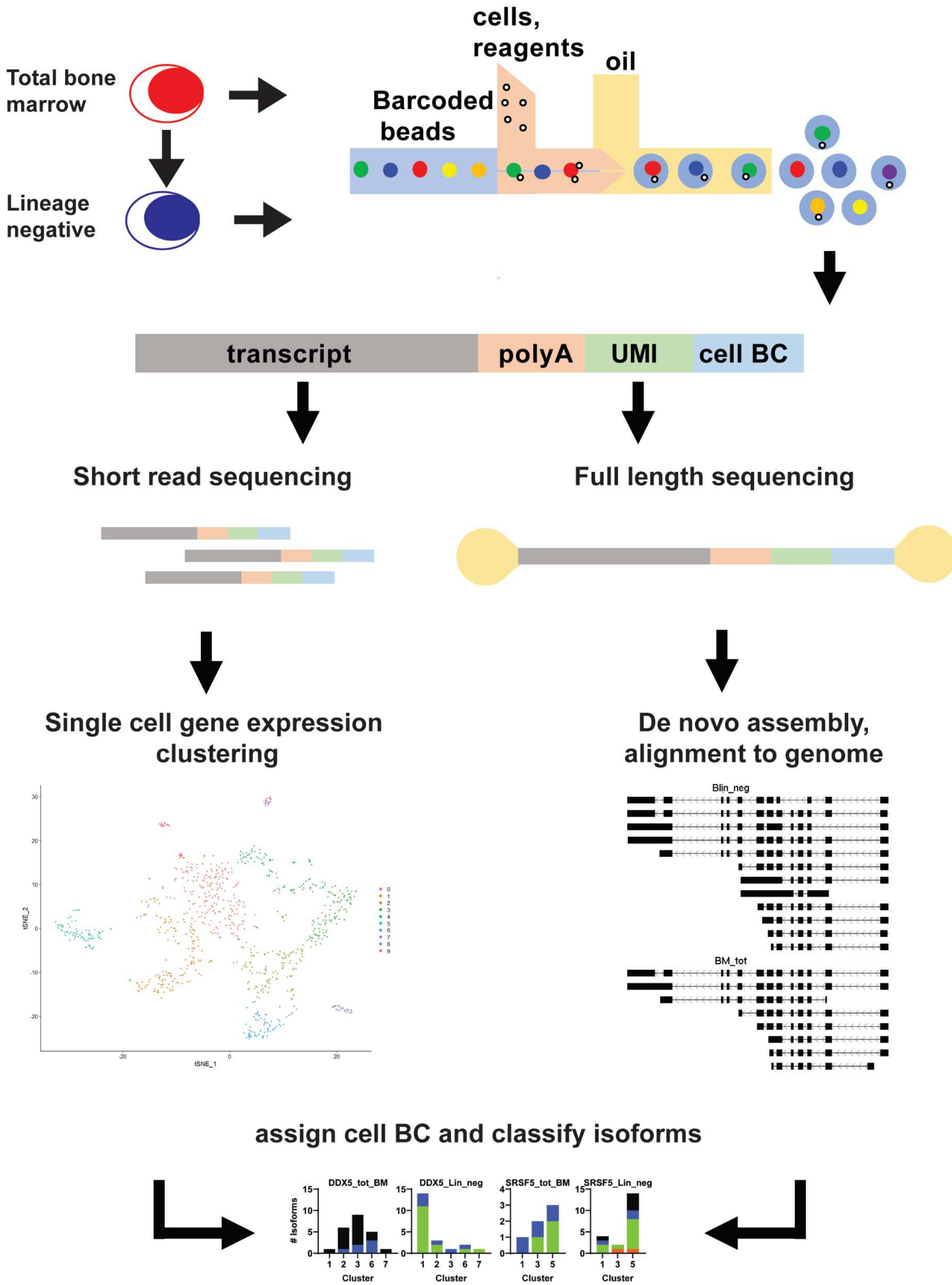


Figure 1

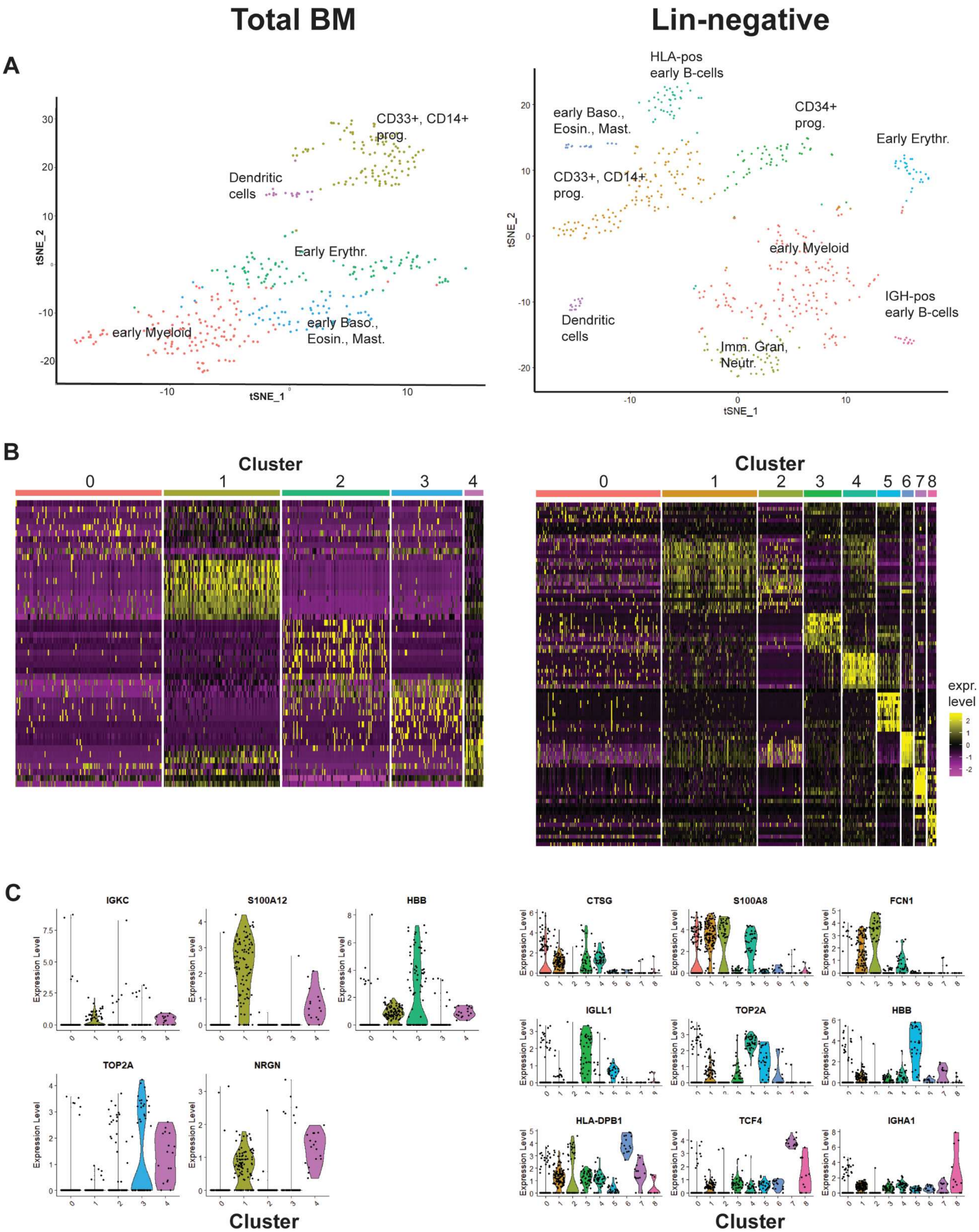
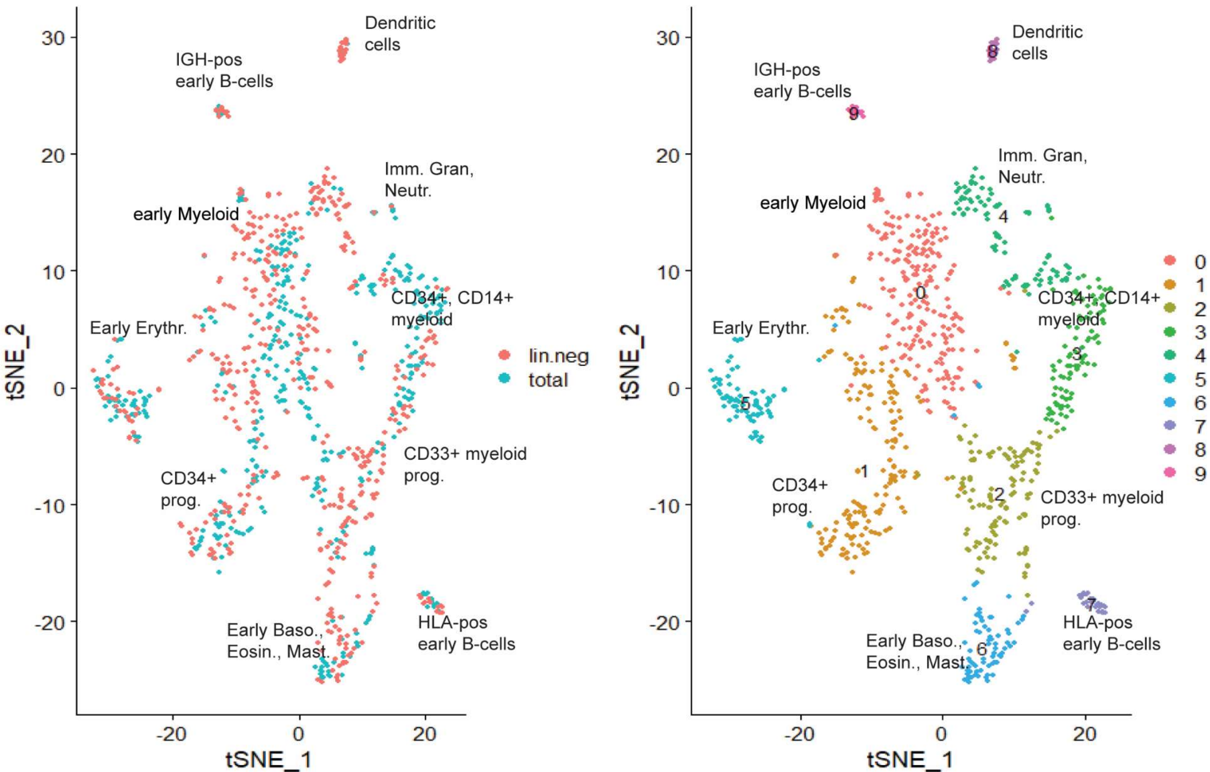

Total BM
Lin-negative
A
CD33+, CD14+ prog.
Dendritic cells
Early Erythr.
early Myeloid
early Baso., Eosin., Mast.
tSNE_1
tSNE_2
HLA-pos early B-cells
CD34+ prog.
Early Erythr.
early Baso., Eosin., Mast.
CD33+, CD14+ prog.
early Myeloid
IGH-pos early B-cells
Dendritic cells
Imm. Gran, Neutr.
B
Cluster
expr. level
C
IGKC
S100A12
HBB
TOP2A
NRGN
CTSG
S100A8
FCN1
IGLL1
TOP2A
HBB
HLA-DPB1
TCF4
IGHA1
Expression Level
Cluster

Figure 2

A


Dendritic cells
IGH-pos early B-cells
Imm. Gran, Neutr.
early Myeloid
CD34+, CD14+ myeloid
Early Erythr.
CD34+ prog.
CD33+ myeloid prog.
HLA-pos early B-cells
Early Baso., Eosin., Mast.
lin.neg
total
tSNE_1
tSNE_2
0
1
2
3
4
5
6
7
8
9

B

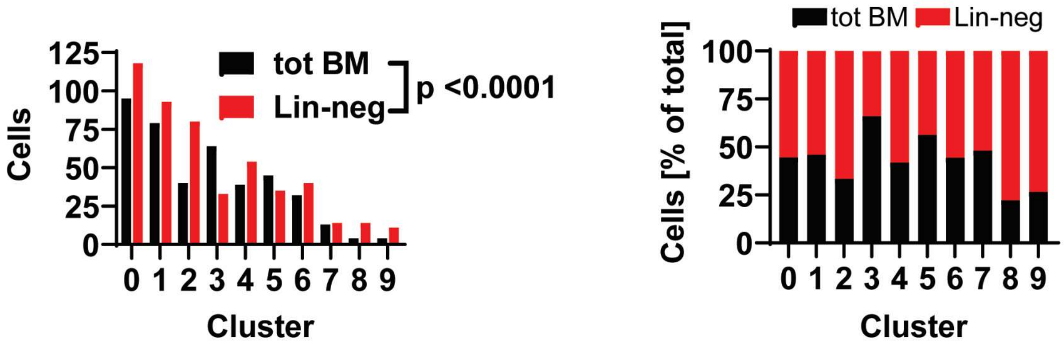

tot BM
Lin-neg
p <0.0001
Cells
Cluster
Cells [% of total]
0 1 2 3 4 5 6 7 8 9

Figure 3

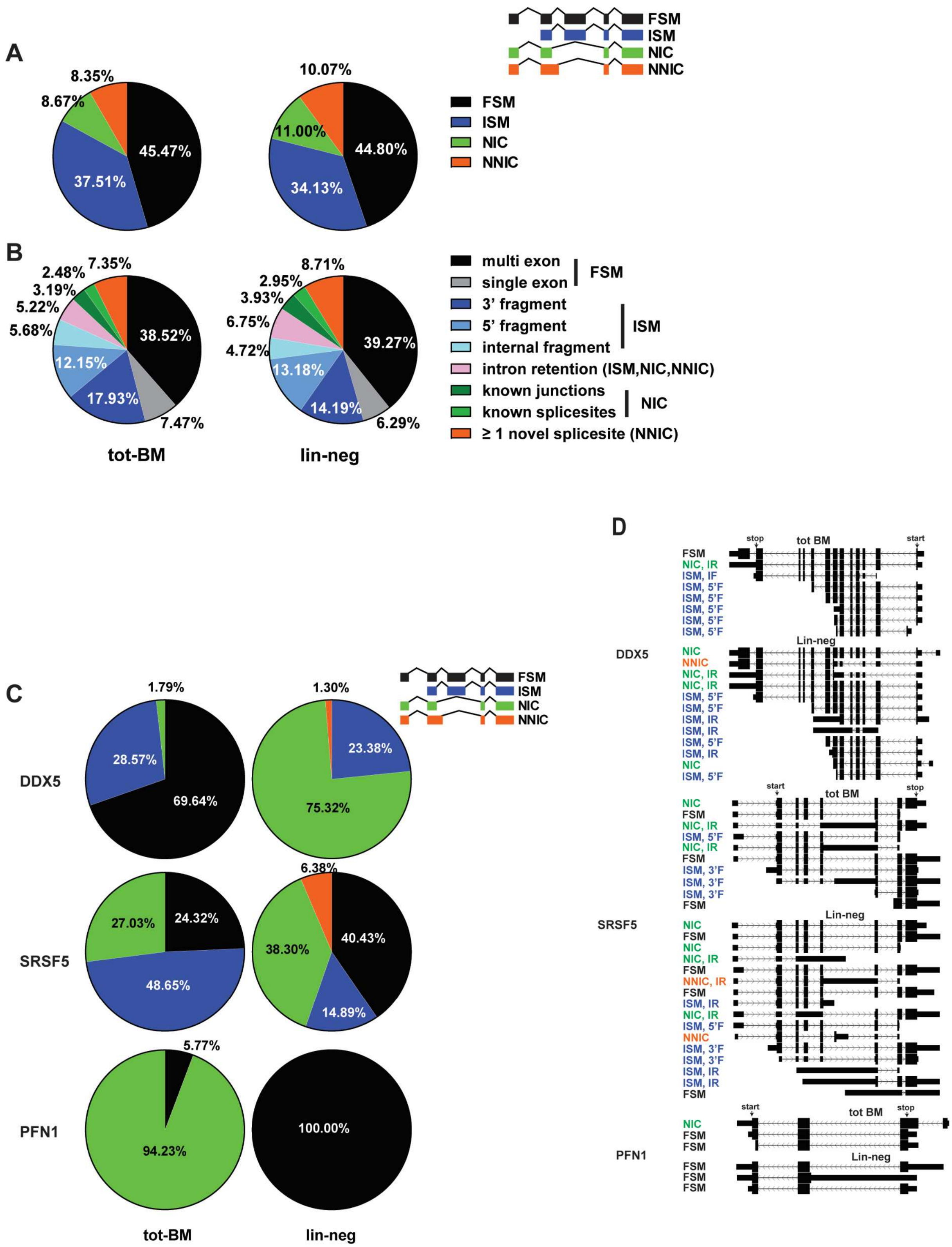

A
FSM
ISM
NIC
NNIC
8.35%
8.67%
45.47%
37.51%
10.07%
11.00%
44.80%
34.13%
B
multi exon
single exon
FSM
3' fragment
5' fragment
internal fragment
ISM
intron retention (ISM,NIC,NNIC)
known junctions
known splicesites
NIC
≥ 1 novel splicesite (NNIC)
2.48%
7.35%
3.19%
5.22%
5.68%
38.52%
12.15%
17.93%
7.47%
8.71%
2.95%
3.93%
6.75%
4.72%
39.27%
13.18%
14.19%
6.29%
tot-BM
lin-neg
C
DDX5
SRSF5
PFN1
1.79%
28.57%
69.64%
1.30%
23.38%
75.32%
27.03%
24.32%
48.65%
6.38%
38.30%
40.43%
14.89%
5.77%
94.23%
100.00%
tot-BM
lin-neg
D
DDX5
SRSF5
PFN1
tot BM
Lin-neg
start
stop

Figure 4

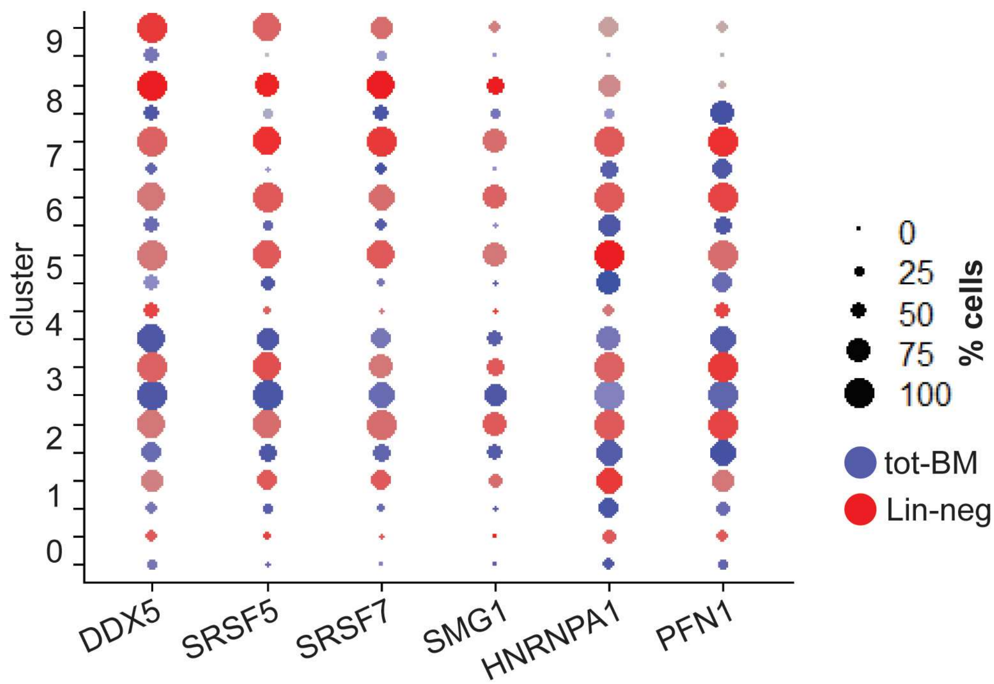

cluster
9
8
7
6
5
4
3
2
1
0
DDX5
SRSF5
SRSF7
SMG1
HNRNPA1
PFN1
0
25
50
75
100
% cells
tot-BM
Lin-neg

Figure 5

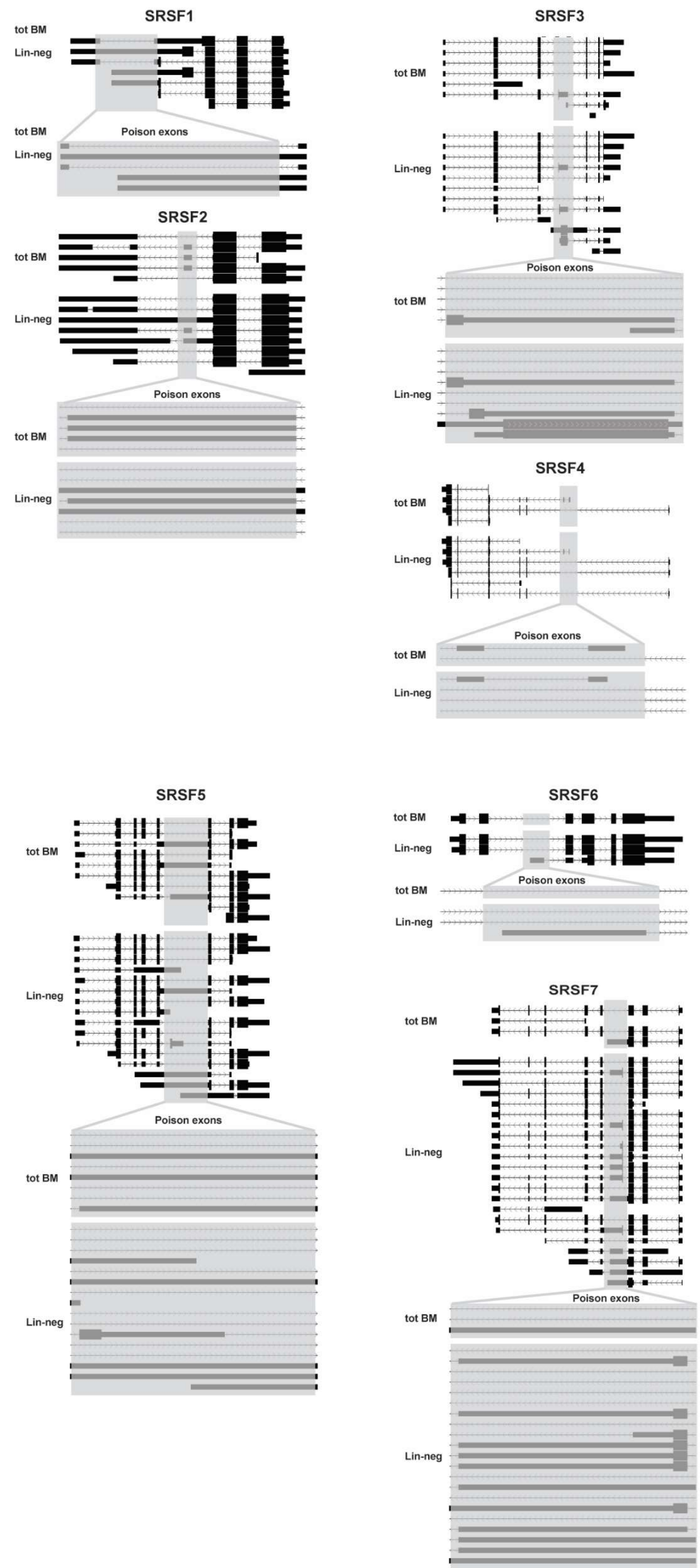

SRSF1
tot BM
Lin-neg
Poison exons
SRSF2
SRSF3
SRSF4
SRSF5
SRSF6
SRSF7

Figure 6

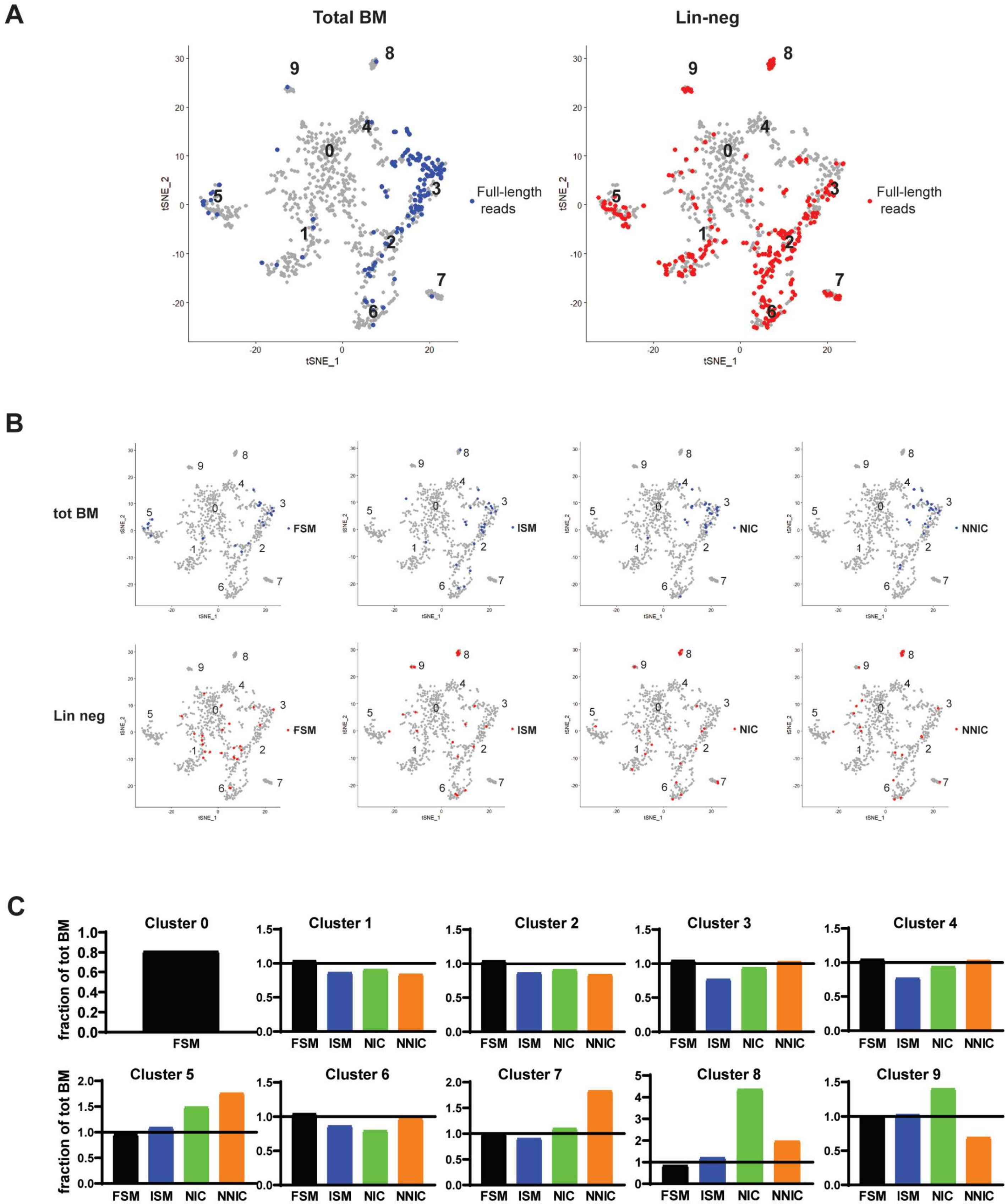

A
Total BM
Lin-neg
Full-length reads
tSNE_1
tSNE_2
B
tot BM
Lin neg
FSM
ISM
NIC
NNIC
C
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
fraction of tot BM
FSM
ISM
NIC
NNIC

Figure 7

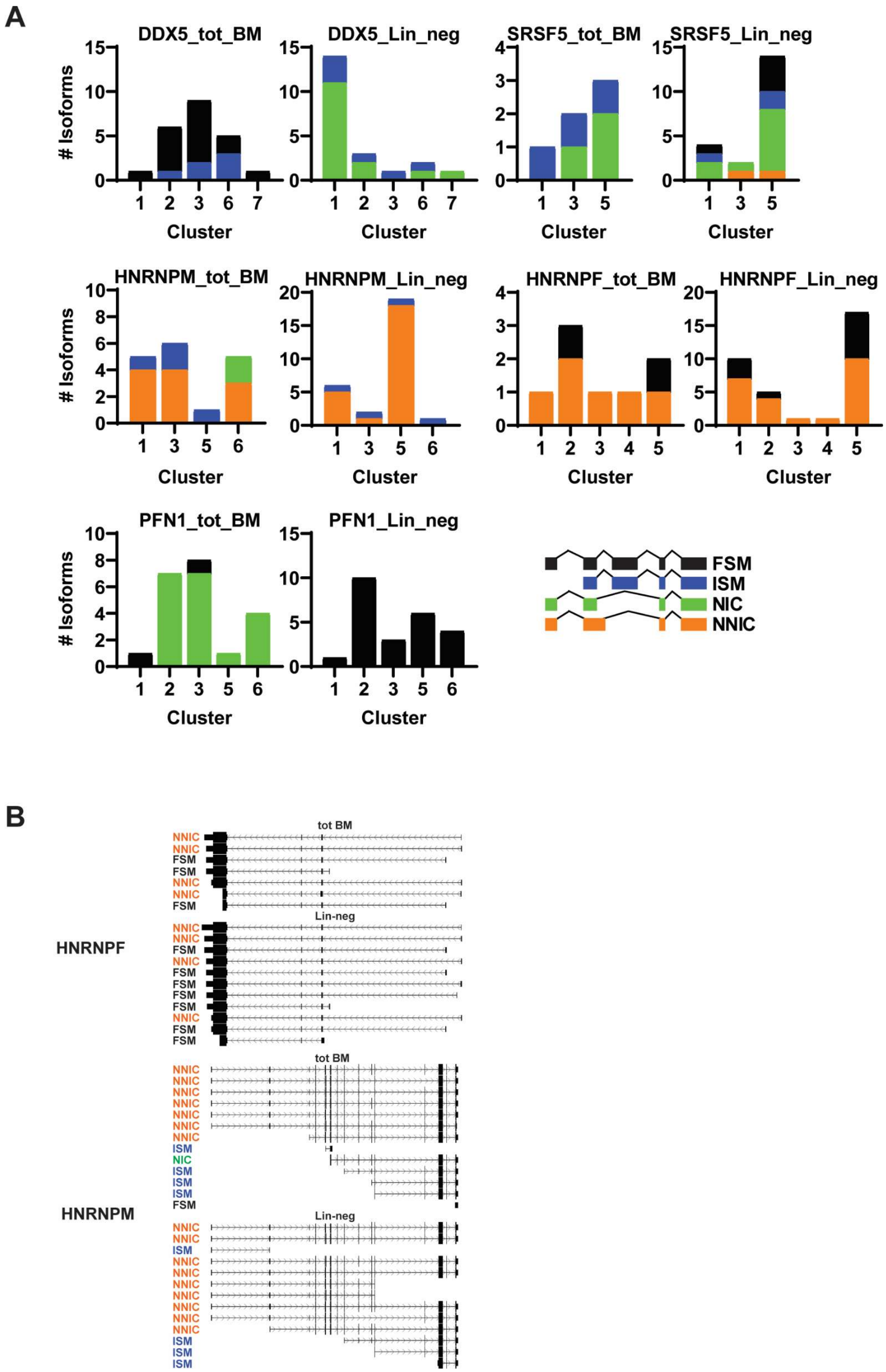
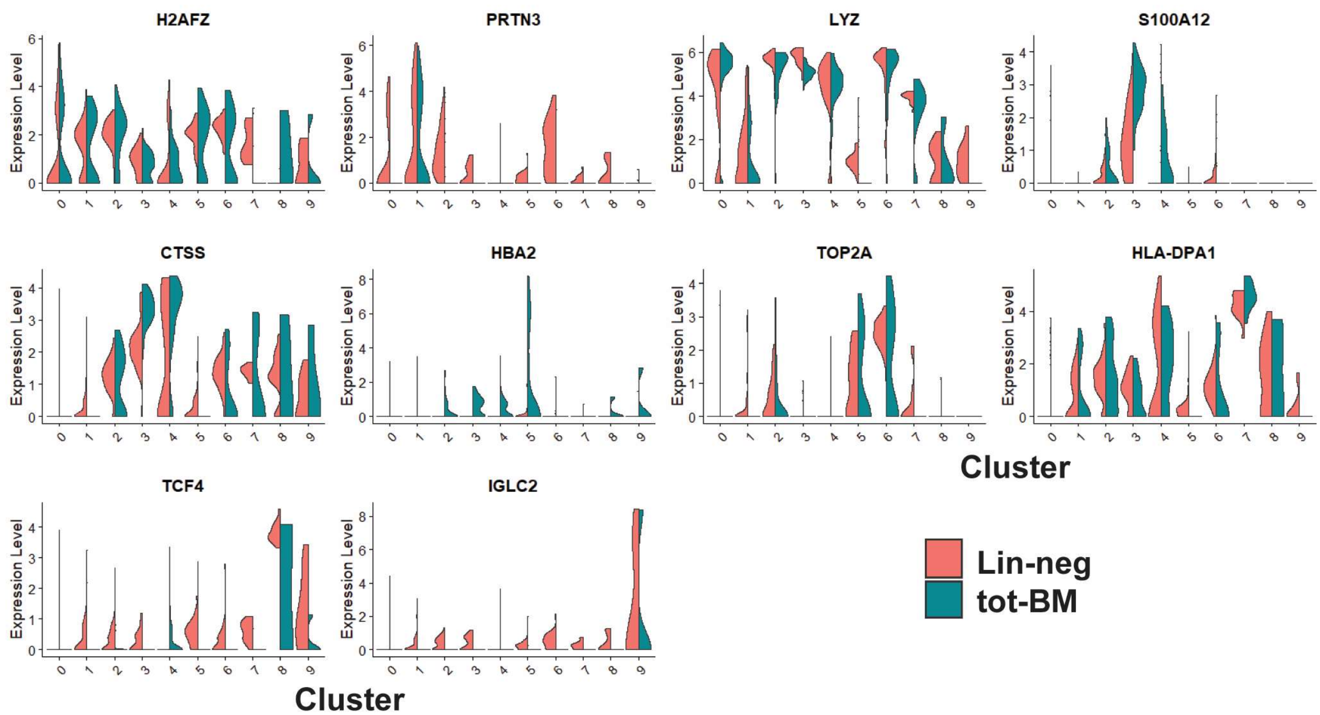

A
DDX5_tot_BM
DDX5_Lin_neg
SRSF5_tot_BM
SRSF5_Lin_neg
HNRNPM_tot_BM
HNRNPM_Lin_neg
HNRNPF_tot_BM
HNRNPF_Lin_neg
PFN1_tot_BM
PFN1_Lin_neg
Isoforms
Cluster
FSM
ISM
NIC
NNIC
B
HNRNPF
HNRNPM
tot BM
Lin-neg

Figure S1



H2AFZ
PRTN3
LYZ
S100A12
CTSS
HBA2
TOP2A
HLA-DPA1
TCF4
IGLC2
Expression Level
Cluster
Lin-neg
tot-BM

Table 1: Total number of genes and mRNA molecules (UMI) from short-read data in single cells

sample	genes	UMI
BM_tot	7,212	57,322
Blin_neg	8,758	133,744

Table 2: Cluster gene type annotation by expression aligned by row (BioGPS, The Human Protein Atlas, Pellin et al. 2019, Velten et al. 2017)

cluster	total BM	cluster	Lineage negative	cluster	combination cluster
0	early Myeloid, Lyz, MPO	0	early Myeloid, Lyz, MPO	0	early Myeloid, Lyz, MPO
		3	CD34+ progenitor	1	CD34+ progenitor
1	CD33+, CD14+ myeloid prog.	1	CD33+, CD14+ myeloid prog.	2	CD33+ myeloid prog.
				3	CD34+, CD14+ myeloid
		2	Imm. Gran, Neutr.	4	Imm. Gran, Neutr.
2	early Megakar, Erythr.	5	early Megakar, Erythr.	5	early Megakar, Erythr.
3	early Baso., Eosin., Mast.	4	early Baso., Eosin., Mast.	6	early Baso., Eosin., Mast.
		6	HLA-pos early B-cells	7	HLA-pos early B-cells
4	Dendritic cells	7	Dendritic cells	8	Dendritic cells
		8	IGH-pos early B-cells	9	IGH-pos early B-cells

Table 3: Number of cells in each cluster after filtering data with Seurat package. The distribution of cells into clusters is significantly different between tot-BM and Lin-neg cell populations (P<0.0001; chi-square test)

cluster	tot-BM	lin.neg	all	
0	95	118	**213**	chi-square test: p< 0.0001
1	79	93	**172**	
2	40	80	**120**	
3	64	33	**97**	
4	39	54	**93**	
5	45	35	**80**	
6	32	40	**72**	
7	13	14	**27**	
8	4	14	**18**	
9	4	11	**15**	
Total	**415**	**492**	**907**	
	chi-square test: p< 0.0001			

Table 4 Full-length sequencing workflow using CCS, lima, isoseq3, minimap2 and SQANTI

	tot-BM	Lin-neg
raw reads	39,270,949	44,507,070
CCS --minPasses 3	393,537	444,826
lima require 5p to 3p adapters	274,879	365,226
clip_out_UMI_cellBC.py	270,545	358,406
isoseq3		
num_reads_flnc	258,783	347,801
num_reads_flnc_polya	256,155	342,805
minimap2		
mapped sequences	256,155	342,805
sqanti "flnc"		
unique genes	18,965	24,110
unique isoforms	66,820	104,728
sqanti_filtered		
unique genes	7,670	9,720
unique isoforms	14,781	23,101

Table 5 Alternative splicing events as categorized by SQANTI are significantly different between tot-BM and Lin-neg cell populations (P<0.0001; chi-square test)

structural_category	subcategory	tot-BM [#]	tot-BM [%]	Lin_neg [#]	Lin_neg [%]	
full-splice_match	multi-exon	5160	38.52%	8063	39.27%	chi-square test: p<0.0001
full-splice_match	single-exon	1001	7.47%	1291	6.29%	
incomplete-splice_match	3prime_fragment	2402	17.93%	2913	14.19%	
incomplete-splice_match	5prime_fragment	1627	12.15%	2706	13.18%	
incomplete-splice_match	internal_fragment	761	5.68%	970	4.72%	
ISM, NIC, NNIC	intron_retention	699	5.22%	1385	6.75%	
novel_in_catalog	combination_of_known_junctions	427	3.19%	807	3.93%	
novel_in_catalog	combination_of_known_splicesites	332	2.48%	606	2.95%	
novel_not_in_catalog	at_least_one_novel_splicesite	985	7.35%	1789	8.71%	

Table 6 Genes with the most novel isoforms in tot-BM and Lin-neg samples (FSM: Full Splice Match, ISM: Incomplete Splice Match, NIC: Novel in Catalog, NNIC: Novel not in Catalog)

ID	name	sample	FSM	ISM	NIC	NNIC	total
BSG	basigin (Ok blood group)	**Lin-neg**	0	0	0	22	22
DDX5	DEAD-box helicase 5	**Lin-neg**	0	18	58	1	77
GAS5	growth arrest specific 5	**tot-BM**	5	0	17	0	22
H2AFY	macroH2A.1 histone	**tot-BM**	12	3	39	1	55
		Lin-neg	1	5	89	1	96
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	**tot-BM**	6	1	0	19	26
HNRNPF	heterogeneous nuclear ribonucleoprotein F	**tot-BM**	6	0	0	24	30
		Lin-neg	20	0	0	38	58
HNRNPK	heterogeneous nuclear ribonucleoprotein K	**Lin-neg**	0	2	1	47	50
HNRNPM	heterogeneous nuclear ribonucleoprotein M	**tot-BM**	1	6	2	44	53
		Lin-neg	0	5	0	56	61
IDH3G	isocitrate dehydrogenase (NAD(+)) 3 non-catalytic subunit gamma	**tot-BM**	0	0	20	0	20
LMO2	LIM domain only 2	**tot-BM**	2	2	0	24	28
		Lin-neg	0	1	15	21	37
OAZ1	ornithine decarboxylase antizyme 1	**tot-BM**	3	4	134	3	144
		Lin-neg	2	0	120	1	123
PFN1	profilin 1	**tot-BM**	3	0	49	0	52
PHB2	prohibitin 2	**Lin-neg**	0	1	0	49	50
PPP1R14B	protein phosphatase 1 regulatory inhibitor subunit 14B	**Lin-neg**	1	0	0	39	40
PRPF40A	pre-mRNA processing factor 40 homolog A	**Lin-neg**	0	3	2	18	23
RNASET2	ribonuclease T2	**Lin-neg**	0	0	2	29	31
SF1	splicing factor 1	**tot-BM**	2	4	13	1	20
SLC7A7	solute carrier family 7 member 7	**tot-BM**	0	2	28	0	30
SRSF5	Serine And Arginine Rich Splicing Factor 5	**Lin-neg**	19	7	18	3	47
TMEM259	transmembrane protein 259	**Lin-neg**	1	1	19	0	21

Table 7
Isoforms per cluster. The distribution of isoforms in the different clusters is significantly different between tot-BM and Lin-neg cell populations (P<0.0001; chi-square test).

cluster	tot-BM isoforms	Lin_neg isoforms	overlapping genes	
0	4	89	0	chi-square test: p< 0.0001
1	3529	11705	949	
2	5805	8799	1149	
3	10023	2022	634	
4	7915	413	211	
5	2059	20915	690	
6	3369	3484	648	
7	961	4872	333	
8	152	662	27	
9	160	1113	34	
total	**33977**	**54074**	**4675**	
	chi-square test: p< 0.0001			

Table S1 Tot-BM marker genes for each cluster. Statistics calculated comparing expression in cluster-specific cells vs cells in all other clusters. Columns are: Log-fold change, multiple test p-value, Bonferroni adjusted p-value, Percentage of positive cells in spec. cluster, Percentage of positive cells in all other clusters,

gene	name	cluster	Fold change (log)	p-value	p-value (adj)	pct. spec custer	pct. other clusters
IGKC	major histocompatibility complex, class II, DM alpha	0	1.7065626	1.71E-06	2.01E-02	0.038	0.240
MPO	myeloperoxidase	0	0.85762519	0.00041	1.00E+00	0.583	0.548
HEXB	hypophosphatemic bone disease	0	0.832369	0.00039	1.00E+00	0.129	0.343
HLA-DMA	hexosaminidase subunit beta	0	0.75546393	0.00554	1.00E+00	0.152	0.329
S100A12	S100 calcium binding protein A12	1	2.45585861	2.25E-64	2.64E-60	0.886	0.058
CTSS	cathepsin S	1	2.37620171	2.63E-61	3.08E-57	1.000	0.174
FCN1	ficolin 1	1	2.06343714	7.21E-61	8.44E-57	0.981	0.132
RGS2	regulator of G-protein signaling 2	1	2.03424637	1.86E-66	2.18E-62	0.943	0.068
HBB	hemoglobin subunit beta	2	2.17273693	8.34E-06	9.76E-02	0.485	0.371
HBD	hemoglobin subunit delta	2	2.00152579	5.75E-12	6.73E-08	0.237	0.028
AHSP	alpha hemoglobin stabilizing protein	2	1.99110243	1.08E-06	1.26E-02	0.247	0.082
PRDX2	peroxiredoxin 2	2	1.7008398	1.56E-10	1.83E-06	0.443	0.186
TOP2A	topoisomerase (DNA) II alpha	3	1.43339091	1.12E-06	1.31E-02	0.359	0.140
CENPF	centromere protein F	3	1.35619246	4.09E-07	4.79E-03	0.359	0.131
MKI67	lysozyme	3	1.31678345	5.33E-07	6.24E-03	0.375	0.145
UBE2C	ubiquitin conjugating enzyme E2 C	3	1.30342572	1.43E-05	1.67E-01	0.219	0.063
NRGN	neurogranin	4	0.81717354	3.61E-14	4.23E-10	1.000	0.226
MKI67	marker of proliferation Ki-67	4	0.81578158	1.64E-12	1.92E-08	0.882	0.151
EREG	epiregulin	4	0.79205986	3.33E-06	3.90E-02	0.706	0.216
LYZ	immunoglobulin kappa constant	4	0.78636841	5.29E-06	6.20E-02	1.000	0.731

Table S2 Lin-neg marker genes for each cluster. . Statistics calculated comparing expression in cluster-specific cells vs cells in all other clusters. Columns are: Log-fold change, multiple test p-value, Bonferroni adjusted p-value, Percentage of positive cells in spec. cluster, Percentage of positive cells in all other clusters,

gene	name	cluster	Fold change (log)	p-value	p-value (adj)	pct. spec custer	pct. other clusters
CTSG	cathepsin G	0	1.242	4.80E-03	1.00E+00	0.285	0.575
DEFA4	defensin alpha 4	0	1.201	3.21E-03	1.00E+00	0.032	0.117
ZNF90	zinc finger protein 90	0	1.182	5.71E-04	1.00E+00	0.032	0.138
HIST3H2A	histone cluster 3 H2A	0	0.941	3.21E-03	1.00E+00	0.032	0.117
S100A8	S100 calcium binding protein A8	1	0.998	3.83E-22	5.43E-18	0.933	0.481
LYZ	lysozyme	1	0.963	6.20E-27	8.79E-23	0.983	0.817
CXCL8	C-X-C motif chemokine ligand 8	1	0.952	5.89E-24	8.36E-20	0.592	0.145
SLC2A3	solute carrier family 2 member 3	1	0.797	4.36E-35	6.19E-31	0.800	0.204
FCN1	ficolin 1	2	1.953	6.61E-08	9.38E-04	0.554	0.328
CTSS	cathepsin S	2	1.719	1.58E-03	1.00E+00	0.536	0.571
S100A9	S100 calcium binding protein A9	2	1.352	1.24E-09	1.76E-05	0.786	0.622
VCAN	versican	2	1.302	1.09E-04	1.00E+00	0.500	0.346
IGLL1	immunoglobulin lambda like polypeptide 1	3	1.415	4.29E-27	6.08E-23	0.830	0.175
SPINK2	serine peptidase inhibitor, Kazal type 2	3	1.103	7.58E-34	1.08E-29	0.745	0.088
C1QTNF4	C1q and tumor necrosis factor related protein 4	3	1.077	1.33E-43	1.88E-39	0.894	0.097
PRSS57	protease, serine 57	3	0.961	1.29E-27	1.83E-23	0.957	0.245
TOP2A	topoisomerase	4	1.554	5.27E-29	7.48E-25	1.000	0.300
CLC	Charcot-Leyden crystal galectin	4	1.405	5.65E-05	8.01E-01	0.190	0.042
MKI67	marker of proliferation Ki-67	4	1.399	6.37E-27	9.04E-23	0.976	0.324
CENPF	centromere protein F	4	1.366	3.65E-24	5.18E-20	0.952	0.298
HBB	hemoglobin subunit beta	5	3.089	2.26E-20	3.21E-16	1.000	0.348
AHSP	alpha hemoglobin stabilizing protein	5	1.940	1.83E-39	2.59E-35	0.897	0.084
PRDX2	peroxiredoxin 2	5	1.886	9.85E-23	1.40E-18	1.000	0.326
HBA1	hemoglobin subunit alpha 1	5	1.840	9.97E-35	1.41E-30	0.690	0.045
HLA-DPB1	major histocompatibility complex, class II, DP beta 1	6	2.567	2.58E-11	3.66E-07	1.000	0.545
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	6	2.432	3.99E-14	5.67E-10	1.000	0.304
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	6	2.412	3.98E-11	5.65E-07	1.000	0.570
ID2	inhibitor of DNA binding 2, HLH protein	6	2.317	3.72E-16	5.27E-12	1.000	0.260
TCF4	transcription factor 4	7	3.115	6.23E-13	8.84E-09	1.000	0.374
GZMB	granzyme B	7	2.712	2.78E-32	3.94E-28	0.929	0.065
HIST1H2BG	histone cluster 1 H2B family member g	7	2.620	1.70E-26	2.42E-22	0.929	0.088
CCDC50	coiled-coil domain containing 50	7	2.594	4.03E-15	5.72E-11	1.000	0.266
IGHA1	immunoglobulin heavy constant alpha 1	8	4.840	4.83E-04	1.00E+00	0.818	0.516
IGKC	immunoglobulin kappa constant	8	4.496	6.32E-05	8.96E-01	1.000	0.655
IGLC3	immunoglobulin lambda constant 3	8	4.474	1.56E-04	1.00E+00	0.364	0.075
IGHG1	immunoglobulin heavy constant gamma 1	8	3.818	3.08E-09	4.37E-05	1.000	0.356

Table S3, Combined tot-BM and Lin-neg marker genes for each cluster. . Statistics calculated comparing expression in cluster-specific cells vs cells in all other clusters. Columns are: Log-fold change, multiple test p-value, Bonferroni adjusted p-value, Percentage of positive cells in spec. cluster, Percentage of positive cells in all other clusters,

gene	name	cluster	Fold change (log)	p-value	p-value (adj)	pct. spec custer	pct. other clusters
EIF5	eukaryotic translation initiation factor 5	0	0.822	8.43E-16	1.21E-11	0.146	0.559
UBB	ubiquitin B	0	0.608	3.21E-07	4.59E-03	0.272	0.657
CFD	complement factor D	0	0.698	7.09E-06	1.02E-01	0.188	0.441
H2AFZ	H2A histone family member Z	0	0.880	8.16E-04	1.00E+00	0.352	0.703
IGLL1	immunoglobulin lambda like polypeptide 1	1	1.632	1.19E-37	1.70E-33	0.465	0.093
PRTN3	proteinase 3	1	1.978	3.47E-11	4.96E-07	0.517	0.333
ELANE	elastase, neutrophil expressed	1	1.794	3.73E-07	5.33E-03	0.535	0.405
CTSG	cathepsin G	1	1.885	1.42E-06	2.04E-02	0.453	0.309
CDCA7	cell division cycle associated 7	2	0.612	1.53E-52	2.19E-48	0.725	0.130
HELLS	helicase, lymphoid-specific	2	0.568	1.96E-43	2.81E-39	0.775	0.183
PCNA	proliferating cell nuclear antigen	2	0.702	2.51E-38	3.60E-34	0.867	0.271
LYZ	lysozyme	2	0.735	8.76E-24	1.25E-19	0.992	0.776
S100A12	S100 calcium binding protein A12	3	2.216	4.21E-82	6.02E-78	0.928	0.143
SLC2A3	solute carrier family 2 member 3	3	1.847	3.47E-67	4.97E-63	0.990	0.259
S100A8	S100 calcium binding protein A8	3	1.816	8.41E-51	1.20E-46	1.000	0.498
S100A9	S100 calcium binding protein A9	3	1.813	1.05E-50	1.50E-46	1.000	0.556
FCN1	ficolin 1	4	1.387	1.09E-21	1.56E-17	0.731	0.307
CTSS	cathepsin S	4	1.525	1.59E-14	2.28E-10	0.677	0.461
HLA-DQB1	major histocompatibility complex, class II, DQ beta 1	4	1.298	9.35E-14	1.34E-09	0.710	0.471
POU2F2	POU class 2 homeobox 2	4	1.314	2.48E-08	3.54E-04	0.473	0.290
AHSP	alpha hemoglobin stabilizing protein	5	2.410	4.38E-55	6.26E-51	0.650	0.076
HBB	hemoglobin subunit beta	5	4.129	1.52E-38	2.17E-34	0.875	0.345
HBA1	hemoglobin subunit alpha 1	5	4.033	2.44E-21	3.49E-17	0.525	0.139
HBA2	hemoglobin subunit alpha 2	5	4.336	2.40E-06	3.43E-02	0.300	0.127
CENPE	centromere protein E	6	1.399	3.07E-39	4.40E-35	0.708	0.132
MKI67	marker of proliferation Ki-67	6	1.447	1.93E-34	2.77E-30	0.819	0.243
TOP2A	topoisomerase	6	1.582	3.49E-32	4.99E-28	0.792	0.230
CENPF	centromere protein F	6	1.342	1.33E-28	1.90E-24	0.764	0.225
HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1	7	2.597	1.98E-22	2.83E-18	0.889	0.227
HLA-DPA1	major histocompatibility complex, class II, DP alpha 1	7	2.663	2.05E-20	2.93E-16	1.000	0.490
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	7	2.246	7.19E-19	1.03E-14	1.000	0.555
HLA-DPB1	major histocompatibility complex, class II, DP beta 1	7	2.524	1.21E-16	1.73E-12	0.926	0.452
IRF4	interferon regulatory factor 4	8	2.474	1.07E-49	1.53E-45	0.778	0.035
HIST1H2BG	histone cluster 1 H2B family member g	8	2.448	4.12E-30	5.89E-26	0.722	0.058
CCDC50	coiled-coil domain containing 50	8	2.615	1.67E-18	2.39E-14	0.889	0.196
TCF4	transcription factor 4	8	3.093	4.06E-15	5.81E-11	0.889	0.268
IGHG1	immunoglobulin heavy constant gamma 1	9	4.516	1.37E-15	1.96E-11	0.933	0.211
IGKC	immunoglobulin kappa constant	9	5.356	2.46E-08	3.52E-04	0.933	0.432
IGHA1	immunoglobulin heavy constant alpha 1	9	5.013	1.63E-06	2.33E-02	0.733	0.280
IGLC2	immunoglobulin lambda constant 2	9	6.099	2.43E-03	1.00E+00	0.467	0.215