Singletrome: A method to analyze and enhance the transcriptome with long noncoding RNAs for single cell analysis

Raza Ur Rahman[1,2], Iftikhar Ahmad[3], Robert Sparks[1], Amel Ben Saad[1], Alan Mullen[1,2]

[1]Division of Gastroenterology, UMass Chan Medical School, Worcester, MA, USA
[2]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[3]Department of Computer Science and Information Technology, University of Engineering and Technology, Peshawar, Pakistan

Abstract

Single cell RNA sequencing (scRNA-seq) has revolutionized the study of gene expression in individual cell types from heterogeneous tissue. To date, scRNA-seq studies have focused primarily on expression of protein-coding genes, as the functions of these genes are more broadly understood and more readily linked to phenotype. However, long noncoding RNAs (lncRNAs) are even more diverse than protein-coding genes, yet remain an underexplored component of scRNA-seq data. While less is known about lncRNAs, they are widely expressed and regulate cell development and the progression of diseases including cancer and liver disease. Dedicated lncRNA annotation databases continue to expand, but most lncRNA genes are not yet included in reference annotations applied to scRNA-seq analysis. Simply creating a new annotation containing known protein-coding and lncRNA genes is not sufficient, because the addition of lncRNA genes that overlap in sense and antisense with protein-coding genes will affect how reads are counted for both protein-coding and lncRNA genes. Here we introduce Singletrome, an enhanced human lncRNA genome annotation for scRNA-seq analysis, by merging protein-coding and lncRNA databases with additional filters for quality control. Using Singletrome to characterize expression of lncRNAs in human peripheral blood mononuclear cell (PBMC) and liver scRNA-seq samples, we observed an increase in the number of reads mapped to exons, detected more lncRNA genes, and observed a decrease in uniquely mapped transcriptome reads, indicating improved mapping specificity. Moreover, we were able to cluster cell types based solely on lncRNAs expression, providing evidence of the depth and diversity of lncRNA reads contained in scRNA-seq data. Our analysis identified lncRNAs differentially expressed in specific cell types with development of liver fibrosis. Importantly, lncRNAs alone were able to predict cell types and human disease pathology through the application of machine learning. This comprehensive annotation will allow mapping of lncRNA expression across cell types of the human body facilitating the development of an atlas of human lncRNAs in health and disease.

Introduction

Long noncoding (lncRNAs) comprise a diverse class of transcripts that regulate pathology, including cancer (Huang et al. 2017), immunity (Kotzin et al. 2016), and liver disease (Mahpour and Mullen 2021). LncRNA transcripts are at least 200 nucleotides in length, 5' capped, 3' polyadenylated, and are not known to code for proteins (Statello et al. 2021). The functions of individual lncRNAs are diverse, with new activities described as additional lncRNAs are investigated. For example, *Evx1as* regulates mesendoderm differentiation through cis regulation of Even-skipped homeobox 1 (*Evx1*) (Bell et al. 2016). *DIGIT* (*GSC-DT*) interacts with BRD3 to control definitive endoderm differentiation (Daneshvar et al. 2020)*,* and *Morrbid* controls the lifespan of immune cells by regulating the transcription of the apoptotic gene *Bcl2l11* through the enrichment of the PRC2 complex at its promoter (Kotzin et al. 2016).

Many lncRNAs also exhibit cell-type-specific patterns of expression (Liu et al. 2016). For example, *LOC646329* is enriched in single radial glia of the human neocortex (Liu et al. 2016), and *Lnc18q22.2* is induced only in hepatocytes in the setting of nonalcoholic steatohepatitis (NASH) (Atanasovska et al. 2017). Cell-type-specific expression patterns observed for many lncRNAs suggest that lncRNA expression could support distinct clustering of cell types in single cell data.

Despite advances in our understanding of the functions of many lncRNAs and frequent examples of cell-type-specific expression, lncRNA discovery is still at a preliminary stage, and there is not yet consensus on the number of lncRNAs in the human genome. GENCODE (v32), the most widely applied genome annotation for human scRNA-seq analysis, contains 16,849 lncRNA genes (Frankish et al. 2019), but databases such as LncExpDB and Noncode now report over 100,000 human lncRNA genes (Li et al. 2021; Fang et al. 2018).

Increasing the number of lncRNAs identified in single cell data cannot be achieved by simply creating new annotations that contain known protein-coding and lncRNA genes, because the addition of tens of thousands of new genes will affect how gene expression is quantified. Current pipelines such as Cell Ranger (Zheng et al. 2017) exclude reads mapping to exons that overlap on the same strand, therefore expanding the number of annotated lncRNA exons may lead to exclusion of additional reads from an increased number of overlapping exons. Furthermore, the assignment of reads that align to antisense transcripts is challenging in part because library preparation artifacts can generate antisense reads at low frequency. For example, the widely used dUTP protocol for stranded RNA-seq (Parkhomchuk et al. 2009) can generate spurious antisense reads ranging from 0.6-3% of the sense signal (Zeng and Mortazavi 2012; Jiang et al. 2011). (Mourão et al. 2019) analyzed 199 strand-specific RNA-seq datasets and discovered that spurious antisense reads are detected in these experiments at levels greater than 1% of sense gene expression levels. Additionally, mis-priming of internal poly-A tracts on RNA or template switching into the poly-T linker have been proposed as possible sources of intronic and antisense reads in single cell gene expression data (Ding et al. 2020). Ultimately, full-length RNA molecule sequencing will help to define authentic antisense RNAs. However, reverse transcriptase-based approaches are predominantly used for sequencing, and special attention needs to be directed towards distinguishing authentic antisense lncRNAs from experimental artifacts, as lncRNAs tend to be expressed at ~10-fold lower levels than protein-coding genes (Cabili et al. 2011; Derrien et al. 2012). It is crucial to develop an approach to minimize the possibility of interpreting the presence of reads antisense to a protein-coding exon as evidence of lncRNA expression if reads are the product of library preparation.

No systematic efforts have been made to analyze all annotated lncRNAs in scRNA-seq data. The most widely used genome annotation for scRNA-seq analysis is GENCODE, which contains only a fraction of annotated lncRNAs in the human genome. Here we develop Singletrome, a comprehensive genome annotation of 110,599 genes consisting of 19,384 protein-coding genes from GENCODE and 91,215 lncRNA genes from LncExpDB, which takes into account the sense and antisense relationship between lncRNAs and protein-coding genes and the distribution of reads across lncRNA transcripts in each dataset. Here we apply Singletrome to analyze single cell data from PBMCs and for analysis of healthy and diseased liver.

Results

Expanding lncRNA annotations in single cell analysis

In order to enhance the current genome annotation for lncRNAs in single cell analysis, we first evaluated how the integration of LncExpDB into GENCODE impacts the annotation. We identified 6309 protein-coding genes (42,868 exons) that overlap on the sense strand with 7531 lncRNA genes (24,357 exons) (Fig 1A & Table 1). We next evaluated lncRNA genes annotated antisense to protein-coding genes and found 10,492 protein-coding genes (47,057 exons) overlap on the antisense strand with 14,212 lncRNA genes (44,062 exons) (Fig 1A & Table 1). This situation is not unique to our new annotation, as 619 protein-coding genes (3514 exons) overlap on the sense strand with 516 lncRNA genes (2106 exons) and 3590 protein-coding genes (12,941 exons) overlap on the antisense strand with 3791 lncRNA genes (8809 exons) in GENCODE (Table 2). We removed the 7531 lncRNA genes from LncExpDB that overlap protein-coding genes on the sense strand (Fig 1B), as it is challenging to prove these lncRNA genes are not isoforms of the protein-coding genes or have coding potential. As a result, reads mapped to the protein-coding exons that overlap on the same strand with these lncRNAs are included to define Unique Molecular Identifier (UMI) counts for the protein-coding genes. To distinguish authentic antisense lncRNAs from potential artifacts, we developed a trimmed lncRNA genome annotation (TLGA) to retain all the non-overlapping lncRNA exonic regions (Fig. 1C). The approach to only count reads mapped to regions of lncRNAs that are not antisense to protein-coding genes reduces the risk of incorrectly calling an lncRNA as expressed based only on antisense reads that might have been generated during library preparation.

In brief, we trimmed lncRNA exons that overlap with protein-coding exons on the opposite strand and also removed an additional flanking 100 nucleotides (nt). We retained exons that were at least 200 nt in length after trimming. Using this strategy, we were able to retain 11,673 of the 14,212 lncRNA genes that contain regions antisense to protein-coding genes. We deleted 2539 lncRNA genes where no exons satisfy the aforementioned criteria. Following these trimming steps we retained 91,215 of 101,285 lncRNA genes (Fig. 1C & Table 1). We then created a comprehensive genome annotation of 110,599 genes consisting of 19,384 protein-coding genes from GENCODE and 91,215 lncRNA genes containing regions that do not overlap with protein-coding genes and refer to this approach as the trimmed lncRNA genome annotation (TLGA). TLGA increased the wealth of lncRNA exons by 4.93 fold (n=428,298), transcripts by 6.46 fold (n=258,106), and genes by 5.41 fold (n=74,366) compared to GENCODE. The inclusion of these additional lncRNA genes may also slightly reduce the total number of uniquely mapped reads, as some reads uniquely mapped in GENCODE will no longer be uniquely mapped with TLGA. (Fig 1D).

Maximizing reads mapped to lncRNAs for downstream analysis

TLGA expands the number of annotated lncRNAs but still excludes regions of 11,673 lncRNAs that partially overlap antisense exons of protein-coding genes. Once we define an lncRNA as expressed in a dataset, the antisense reads could provide additional depth to assist in cell clustering and the definition of genes expressed in specific cell types for follow-up studies. In addition, lncRNAs are expressed at lower levels than protein-coding genes (Fig 2A-B and Supplementary Fig 1A-D), and antisense lncRNAs often have functional activity (Faghihi et al. 2010; Yap et al. 2010), so there are benefits to including as much information for these genes as possible once the thresholds for expression are met.

To assess the impact of trimming, we compared TLGA with an untrimmed lncRNA genome annotation (ULGA). In ULGA, we deleted lncRNA genes overlapping protein-coding genes on the sense strand but included all reads for the antisense overlapping lncRNAs. We mapped PBMCs (pbmc_10k_v3 from 10x Genomics), liver set 1 (GSE115469 (MacParland et al. 2018)) and liver set 2 (GSE136103 (Ramachandran et al. 2019)) scRNA-seq data with ULGA and TLGA to assess the output from each annotation.

More than 1,000 lncRNA genes in each dataset are expressed in ULGA but have no expression in TLGA. Of 14,212 antisense overlapping lncRNAs, 1458 lncRNAs in PBMCs are expressed in ULGA but not TLGA, while 1153 and 1841 lncRNAs are expressed in ULGA but not TLGA in liver sets 1 and 2, respectively (Supplementary Table 1). Reads mapped to these lncRNA genes were aligned only to regions antisense to protein-coding exons (+100nt) in ULGA and have the potential to come only from library preparation artifacts. These lncRNAs were removed from down-stream analysis.

Of the 14,212 antisense overlapping lncRNAs, 4921 lncRNAs in PBMCs, 4194 in liver set 1, and 6675 in liver set 2 are expressed in both TLGA and ULGA. For these lncRNAs, TLGA excluded many reads that could support expression of lncRNA genes where there was corroborating evidence for expression from reads that were not antisense to other genes. The median of reads mapped to these lncRNAs is reduced from 174 to 142 in PBMCs, 45 to 38 in liver set 1 and 70 to 57 in liver set 2 for TLGA as compared to ULGA, and the same trends are observed in each cell type (Fig 2C-D and Supplementary Fig 2A-D and Supplementary Table 1). This analysis suggests that TLGA can be used to identify lncRNAs with the highest confidence in expression but reduces the reads associated with lncRNAs containing exons antisense to other genes. On the contrary the ULGA accounts for all the possible reads mapped to lncRNA genes at the cost of potential library preparation artifacts. To this end, we combined both approaches. We utilized TLGA to define expressed lncRNAs and ULGA to account for all the reads mapped to these lncRNAs.

Two pairs of overlapping protein-coding and lncRNAs genes illustrate these scenarios in PBMCs. *SRGAP2-AS1* overlaps *SRGAP2C* in antisense (Fig 2E). The reads supporting *SRGAP2-AS1* are only within exons antisense to *ARGAP2C* (blue arrows). *SRGAP2-AS1* is defined as not expressed in TLGA and is excluded from further analysis. *HSALNG0137471* is expressed antisense to *DDX3C* (Fig 2F). In this example, there are reads supporting expression of *HSALNG0137471* in regions of exons that are not antisense to *DDX3X* exons (black vertical arrows). This lncRNA is defined as expressed in TLGA. There are additional reads mapped in ULGA closer to the 3' end of *HSALNG0137471* that can be included to provide additional support for expression of this lncRNA.

Read mapping and detected lncRNAs

We analyzed 8.07 billion reads in three publicly available datasets (26 samples) consisting of one PBMC dataset and two liver datasets (Table 3). We mapped all samples to GENCODE, TLGA and ULGA. Genome indices were created using Cell Ranger version 3.1.0 due to its compatibility with different versions of Cell Ranger count (material and methods). The difference in the number of reads mapped to various genomic loci between TLGA and ULGA is minimal (Supplementary Fig 3-4 & Supplementary data 1). However, we observed a significant difference between the reads mapped to GENCODE compared to both ULGA and TLGA. Across the 26 samples, we observed an increase in the reads mapped confidently to exonic regions by 1.46% in ULGA compared to GENCODE, accounting for 118.09 million reads (Supplementary Fig 3B and Supplementary data 1). These results suggest that a fraction of reads not mapped to GENCODE can be uniquely mapped to lncRNAs added with TLGA and ULGA. Furthermore, ULGA captures more lncRNA genes (expressed in at least 10 cells in a dataset) compared to GENCODE (Supplementary Table 2). GENCODE detected 5064, 4800, and 8211 lncRNAs in PBMCs, Liver set 1, and Liver set 2 compared to 25,470, 20,813, and 40,375 in ULGA. In contrast, we observed a decrease in the reads mapped confidently to the genome by 1.19% (96.26 million), intergenic regions by 1.98% (159.93 million), intronic regions by 0.69% (56 million reads), and transcriptome by 0.15% (12.09 million reads) in ULGA compared to GENCODE. The reduced reads in these categories suggest that ULGA and TLGA now captures a small number of reads defined as intergenic or intronic by GENCODE and that a small fraction of reads uniquely mapped in GENCODE are no longer uniquely mapped with the expanded annotation and are discarded (Fig 1D).

Quality control of lncRNA mapping

Many lncRNAs in LncExpDB are not experimentally validated, and we next sought to define additional criteria to support lncRNA gene expression in each dataset. We assessed read distribution across the transcript body to identify lncRNA genes where 1) mapped reads exhibit 5’ bias in 3’ sequenced scRNA-seq libraries and 2) the majority of reads were mapped to a single location in the transcript, as both situations could represent library artifacts or mapping anomalies (Ma and Kingsford 2019). LncRNA genes for which all transcripts met either criteria in a dataset were excluded from further analysis in that dataset.

To obtain the read distribution across the transcript body we utilized RSeQC (Wang, Wang, and Li 2012). RSeQC scales all the transcripts to 100 bins and calculates the number of reads covering each bin position and provides the normalized coverage profile along the gene body. We modified RSeQC to obtain raw read counts (default is normalized read count to 1) for each bin (material and methods). The overall read distribution for lncRNA genes was similar to protein-coding genes in PBMCs (Fig 3A), while liver set 1 and liver set 2 showed more 5’ enrichment than protein-coding genes (Supplementary Fig 5). To assess the read distribution across the transcripts and avoid transcript length bias, we subdivided lncRNAs and protein-coding transcripts based on transcript length (Supplementary Table 3). We observed that lncRNA transcripts from 200-1000 nt in length and greater than 10,000 nt in length have very similar read distribution to protein-coding transcripts (Supplementary Fig 6-8). In contrast, the read distribution for lncRNA transcripts of length more than 1000 nt and less than 10,000 nt exhibit an increase in 5’ enrichment in 3’ sequenced scRNA-seq libraries (Fig 3B-C and Supplementary Fig 6-8).

We next assessed the variability between gene and transcript lengths and found less correlation between gene and transcript length for lncRNAs compared to protein-coding genes (Figure

3D-F & Supplementary Table 4). This finding suggested that the 5' enrichment observed in bulk analysis of lncRNA transcripts could be explained by expression of transcripts of more variable length for a given lncRNA gene, where shorter isoforms could give the appearance of an increased fraction of 5' reads for some lncRNA transcripts. We then evaluated 5' bias for each lncRNA transcript (minimum transcript length 1,000 nt). In total 2445, 3065, and 4486 lncRNA genes had transcripts that were flagged for 5' bias in PBMCs, liver set 1, and liver set 2, respectively (Fig 3G & Supplementary Table 5). Since the observed 5' bias could be explained by more abundant shorter isoforms of an lncRNA, we discarded the lncRNA gene only if all the transcripts were flagged for 5' bias. Using this criteria we discarded 433 lncRNA genes (5685 transcripts) in PBMCs, 488 lncRNA genes (7372 transcripts) in liver set 1, and 928 lncRNA genes (9296 transcripts) in liver set 2 (Supplementary Table 5).

Finally, we evaluated read distribution across lncRNA transcripts to identify potential library artifacts or mapping anomalies. We flagged lncRNA transcripts where reads aligned to one particular region of the full transcript (minimum transcript length 1,000 nt). If the expression of a single bin was greater than the expression of the sum of the remaining 99 bins and this single bin was not in the last 10 bins (denoting the 3' end of the transcript), the transcript was flagged (Fig 3H). In total 606, 644, and 1084 lncRNA genes had transcripts that were flagged in the PBMCs, liver set 1, and liver set 2, respectively. We performed this analysis for all transcripts and discarded the lncRNA gene if all transcripts for a gene displayed this phenomenon. Using these criteria we discarded 67 lncRNA genes (1455 transcripts) in PBMCs, 45 lncRNA genes (1312 transcripts) in liver set 1, and 98 lncRNA genes (2271 transcripts) in liver set 2 (Supplementary Table 5).

After applying these quality control steps, we were able to retain the expression of 23,510, 19,126, and 37,507 high quality lncRNA genes in PBMCs, liver set 1, and liver set 2 (Supplementary Table 5). These lncRNAs were used for all the down-stream analyses.

lncRNAs alone predict most clusters and cell types in single cell data

LncRNA expression can be cell-type-specific (Liu et al. 2016). We applied our new annotation to determine if we could cluster cell types based on lncRNA expression alone. We returned to scRNA-seq data for human PBMCs and liver (Table 3). We mapped scRNA-seq data using Cell Ranger (v6.0.2), and the labels for each cell were retained from the original publications. We clustered cells using data aligned to GENCODE and to Singletrome (with the previously-established filters). Despite lower expression of lncRNAs compared to protein-coding genes (Fig 2A-B & Supplementary Fig 1A-D), we created similar cell clusters based on lncRNAs alone for both PBMCs (Fig 4A-D) and liver (Supplementary Fig 9A-D & Supplementary Fig 10A-D). Clustering by lncRNAs alone showed similar results to GENCODE for most cell clusters but did shift relationships between some clusters and cell types. In PBMCs, we observed that CD4 naive and CD4 memory cells clustered more closely to CD8 naive and CD8 effector cells with lncRNAs alone compared to data aligned to GENCODE (Fig 4A and 4D). In liver set 1, we observed that hepatocytes_5 clustered closely with other hepatocytes with lncRNAs alone as compared to genome annotations containing only protein-coding genes (Supplementary Fig 9A-D). We were not able to separate sub-clusters of hepatocytes (1, 3, 6, and 15) by UMAP using lncRNAs, and these sub-clusters group closely in the original GENCODE annotation (Supplementary Fig 9A). In another example, liver sinusoidal endothelial cells (LSECs)_13 are clustered more closely with LSECs_11 and LSECs_12 with lncRNAs alone (Supplementary Fig 9D) as compared to analysis with protein-coding genes (Supplementary Fig 9A-C). For some cell types, lncRNA alone could not separate populations. For example, gamma-delta (gd)T cells_18, gd T cells_9, alpha-beta (ab) T cells, and natural killer (NK) cells could not be separated by lncRNA-only annotations (Supplementary Fig 9). It is possible that some cell types may have less diversity of lncRNA expression or produce lower levels of lncRNAs transcripts, which could reduce the ability to cluster some cell types by lncRNAs alone.

Since lncRNAs can cluster the majority of cells by cell type, we next aimed to generate an lncRNA-based cell type marker map. We identified marker genes for each cell type relative to all other cell types based on lncRNAs and protein-coding genes in PBMCs (Fig 4E-F) and liver (Supplementary Fig 11-14). While lncRNAs are expressed at lower levels compared to protein-coding genes in all datasets (Fig 4G, Supplementary Fig 15), we were still able to identify lncRNA-based cell markers for PBMCs and liver (Supplementary data 2-4).

Clustering algorithms make assumptions about data distribution. We next trained a machine learner to determine how well lncRNAs can define cell types without the underlying statistical assumptions that are applied to clustering. In order to establish a baseline for comparing cell type predictions, we performed cell type prediction using protein-coding genes and Singletrome (containing all the protein-coding genes and quality filtered lncRNAs). We trained a gradient-boosted decision tree based classifier XGBoost (Extreme Gradient Boosting) on the expression data of protein-coding genes, lncRNAs, and the combination of both from Singletrome (material and methods). Cell type labels were retained from the original publications for PBMCs (13 cell types), liver set 1 (20 cell types), and liver set 2 (12 cell types).

We found that the overall accuracy for predicting cell types using lncRNAs was comparable to that of protein-coding genes for PBMCs (96.39% for protein-coding genes and 90.30% for lncRNAs) and liver set 2 (99.10% for protein-coding genes and 95.43% for lncRNAs) (Fig 4H, Supplementary Fig 16-17 and Supplementary data 5-6). However, liver set 1 had an accuracy of 75.48% for lncRNAs, which is considerably less than the accuracy of 93.66% for protein-coding genes (Supplementary Fig 18 Supplementary data 7). Liver set 1 splits single cell type into multiple clusters based on marker genes from GENCODE, for example it contains six cell clusters of hepatocytes, three clusters of liver sinusoidal endothelial cells (LSECs) and two cell clusters of each macrophages and gd T cells. Five out of six sub-clusters of hepatocytes are closely-associated by UMAP using the original GENCODE annotation (Supplementary Fig 9A), and two sub-clusters of each macrophages and LSECs also cluster closely using the original GENCODE annotation (Supplementary Fig 9A).

To assess the accuracy of predicting cell types rather than sub-clusters of cell types, we merged clusters within the same cell type, retaining 11 cell types in liver set 1. We were able to predict cell types with an accuracy of 98.16% using protein-coding genes, 90.40% using lncRNAs, and 98.16% using Singletrome (Supplementary Fig 19 and Supplementary data 7). These results suggest that lncRNA expression can be used to predict cell types with a similar accuracy to protein-coding genes, even though the original clustering was determined primarily by protein-coding genes.

Long noncoding RNAs in liver fibrosis

To understand the role of lncRNAs in disease, we next analyzed scRNA-seq data of healthy and cirrhotic human liver (liver set 2, GSE136103, (Ramachandran et al. 2019). We again used cell labels from the original study and clustered the cells based on the condition (healthy and cirrhotic) using Singletrome (Supplementary Fig 20), only protein-coding genes from Singletrome (Supplementary Fig 21), and only lncRNAs from Singletrome (Fig 5A). Being able to produce similar clusters of healthy and diseased liver cell types enabled us to perform differential expression analysis of lncRNAs in healthy and cirrhotic liver by cell type.

We detected 937 differentially expressed lncRNA genes (495 up-regulated and 442 down-regulated) between healthy and cirrhotic liver (Supplementary data 8) in cell types including mesenchymal cells, hepatocytes, cholangiocytes, endothelial cells, B cells, plasma B cells, dendritic cells (DCs), mononuclear phagocyte (MPs), innate lymphoid cells (ILCs) and T cells (padj < 0.1 and log2FC > 0.25, Supplementary data 8). We were not able to detect statistically significant differentially regulated lncRNAs in mast cells, and there were not enough mesothelial cells to perform differential expression analysis (Fig 5B) .

LncRNAs induced with cirrhosis include *XIST* and *H19* (Figure 5C-F, Supplementary data 8), which have been shown to promote fibrosis (Wu et al. 2022; Xiao et al. 2019)*. XIST* was identified in the top two induced lncRNAs in cholangiocytes, hepatocytes, and pDCs, while *H19* was identified in top two induced lncRNAs in mesenchymal cells (Fig 5C). Additional lncRNAs that are not yet well-characterized were also identified as differentially expressed between cell types in healthy and cirrhotic liver (Fig 5C-F). These results show that there is sufficient read depth to identify differentially expressed lncRNAs from single cell liver datasets that could have a role in disease.

lncRNAs alone can cluster and predict cell types (Fig 4D,4H and Supplementary Fig 9D, 10D), and we were able to observe the differential expression of lncRNAs linked to cirrhosis in particular cell types. Liver pathology may be influenced by the expression of lncRNAs in affected cell types. To assess whether liver pathology follows observable rules, we trained a machine learner on lncRNA expression data of liver set 2 (GSE136103, (Ramachandran et al. 2019)) and predicted the condition (healthy or cirrhotic) of the affected target cell. Condition (healthy or cirrhotic) annotation for each cell was retained from liver set 2. We applied the XGBoost algorithm to classify healthy and cirrhotic cell types of liver using lncRNAs (material and methods). Based on lncRNA expression alone, the condition of cell types can be predicted with an accuracy of 93.68%, a precision of 93.56%, and a recall of 93.49% (Fig 5G, Supplementary data 6). In order to verify lncRNA based predictions, we trained a separate XGBoost model on the expression data of protein-coding genes, and we were able to predict the condition of the cells with an accuracy of 98.27%, a precision of 98.24%, and a recall of 98.22% (Supplementary data 6). Additionally, Singletrome was able to classify healthy and cirrhotic cells with an accuracy of 98.96%. These results suggest that it is possible for both cell type and disease pathogenicity in single cell data to be reliably predicted through analysis of lncRNA expression alone.

Discussion

We developed Singletrome to interrogate lncRNAs in scRNA-seq data using a custom genome annotation of 110,599 genes consisting of 19,384 protein-coding genes from GENCODE and 91,215 lncRNA genes from LncExpDB (Table 1). We increased the current human lncRNA annotation by greater than five fold and benchmarked the utility of Singletrome by analyzing three publicly available 10x scRNA-seq datasets (Table 3). Mapping metrics such as reads confidently mapped to (i) exonic regions, (ii) genome, (iii) transcriptome, (iv) intronic regions, and (v) total genes detected depend on the expression of genes in each sample and can increase (with the expression of additional lncRNAs) or decrease (due to multi-mapping of reads that were originally confidently-mapped in GENCODE) with the additional lncRNA genes (Supplementary Fig 3-4 & Supplementary data 1). In most of the samples, we observed an increase in the total number of genes detected and reads confidently-mapped to exonic regions, while a decrease in the reads mapped confidently to the genome, transcriptome, intronic regions, and intergenic regions was observed in TLGA and ULGA compared to GENCODE (Supplementary Fig 4 & Supplementary data 1). The balance of reads gained by new lncRNA exons and lost by multi-mapping as a result of these exons can also fluctuate across individual samples. In the example of reads confidently mapped to exon regions (Supplementary Fig 3B), all samples show an increase for TLGA and UGLA except liver-13 and liver 25, where the loss of reads due to multi-mapping outweighs the gain of new exons.

We utilized trimmed lncRNA genome annotation (TLGA) to avoid counting spurious antisense reads when defining lncRNAs that are expressed in a dataset. We then utilized an untrimmed lncRNA genome annotation (ULGA) to account for all the reads mapped to lncRNAs that are defined as expressed by the TLGA (Supplementary Table 1). Using the established quality control filters, we were able to identify lncRNA genes where 1) mapped reads exhibit 5' bias in 3' sequenced scRNA-seq libraries and 2) the majority of reads were mapped to a single location in the transcript, as both situations could represent library artifacts or mapping anomalies(Ma and Kingsford 2019) (Fig 3G-H and Supplementary Table 5).

LncRNA expression can be cell-type-specific (Liu et al. 2016), and we found that most cell types can be clustered by lncRNAs alone (Fig 4A-D and Supplementary Fig 9-10). Clustering by lncRNAs alone was associated with less separation of clusters compared to GENCODE or lncRNAs plus protein-coding genes (Singletrome) as evidenced by the closer proximity of some clusters from the same cell type. These observations may be influenced by lower levels of lncRNA expression or more similarities in expression at the lncRNA level across similar cell types. The additional reads mapped with Singletrome did not result in the clear identification of new clusters within those clusters defined by mapping to GENCODE. However, for these analyses, cell clusters were assigned based on the originally published GENCODE annotations, and future analyses using Singletrome to perform the original clustering may provide additional resolution compared to GENCODE.

To determine cell types based on lncRNAs without the statistical assumptions, we applied the XGBoost classifier for predicting cell type using only lncRNA expression. In order to establish a baseline for comparing cell type predictions using lncRNAs, we additionally trained the XGBoost classifier on the expression data of protein-coding genes and Singletrome. We trained and predicted cell types for all the three datasets (Table 3). The classification of cells into cell types for each dataset (Table 3) was challenging due to (1) multiple cell types in each dataset and (2) imbalanced datasets (Supplementary data 5-7). There were 13 cell types in PBMCs, 20 in liver set 1, and 12 in liver set 2. Furthermore cell types were not represented equally within the dataset. For example, PBMCs have 52 platelets and 2992 CD14+ monocytes, and liver set 1 contains 37 hepatic stellate cells and over a thousand hepatocyte_1 cells. Similarly liver set 2 has 70 mast cells and over 20,000 T cells. A number of machine learning classifiers can be applied for the cell type prediction problem, such as neural networks, support vector machines, random forest and logistic regression. We selected XGBoost as it is a preferred machine learning technique for classification with imbalanced datasets against the aforementioned set of classifiers. (Hernesniemi et al. 2019; Nishio et al. 2018; Ogunleye and Wang 2020).

The overall accuracy for predicting cell types using lncRNAs was comparable to that of protein-coding genes for PBMCs (96.39% for protein-coding genes and 90.30% for lncRNAs) and liver set 2 (99.10% for protein-coding genes and 95.43% for lncRNAs). However, liver set 1 had an accuracy of 75.48% for lncRNAs, which is considerably less than the accuracy of 93.66% for protein-coding genes. Most of the mis-classifications in liver set 1 were amongst sub-clusters of the same cell type (Supplementary data 7). For example 100 cells of *Hepatocytes_3* were classified correctly, but 72 cells of *Hepatocytes_3* are classified as *Hepatocytes_1.* These results indicate that the cells in each subcluster still contain many of the same lncRNAs. When we collapsed cell sub-clusters into a single cell type, we predicted liver set 1 cell types with an accuracy of 98.16% for protein-coding and 90.40% for lncRNAs (Supplementary data 7). These results suggest that lncRNA expression can be used to predict cell types with a similar accuracy to protein-coding genes. However, the prediction accuracy drops when separating sub-clusters of the same cell type. In these cases it is not yet clear whether lncRNAs are slightly less able to predict cell type, or the difference in prediction between lncRNAs and protein-coding genes reflects an inherent bias towards the protein-coding genes in the original cell type labeling.

The ability to cluster and predict most cell types using lncRNA expression alone, demonstrates the depth and diversity of lncRNA transcripts detected in single cell data. The overall goal of Singletrome is to increase the depth of annotations of single cell data and to define differentially expressed lncRNA genes that may regulate disease processes. Comparing cells from healthy and cirrhotic liver (liver set 2), we were able to identify 937 differentially expressed lncRNAs. *XIST* and *H19* are both linked to liver fibrosis (Wu et al. 2022; Xiao et al. 2019) and were identified in our analysis. This suggests that other lncRNAs with similar patterns of expression (Figure 5C-F and Supplementary data 8) may also have activity in liver fibrosis. Our analysis was based on currently available data for healthy and cirrhotic liver. The data includes five cirrhotic livers with different causes of cirrhosis. As liver datasets expand in the future to include additional replicates with cirrhosis from multiple sources of injury and different stages along disease progression, the statistical power will increase to allow identification of additional differentially expressed lncRNAs across all conditions.

lncRNAs alone can cluster and predict cell types (Fig4D, 4H and Supplementary Fig 9D, 10D, 16-19), and we were able to identify differential regulation of lncRNAs linked to cirrhosis in particular cell types. Machine learning also demonstrated the ability of lncRNAs to predict disease (Fig 5G). These analyses demonstrates that lncRNA expression changes significantly in disease and provides further support to suggest that lncRNAs, in addition to protein-coding genes, can serve as biomarkers and mechanistic drivers of disease (Nath et al. 2019; Delás and Hannon 2017; Bolha, Ravnik-Glavač, and Glavač 2017).

The Human Cell Atlas has now mapped more than a million individual cells across 33 organs of the human body (Suo et al. 2022; Domínguez Conde et al. 2022) The focus of these analyses has understandably been on protein-coding genes. This comprehensive genome annotation optimized for scRNA-seq data can now be applied to existing and future single cell data sets to promote the development of an atlas of human lncRNAs in health and disease.

Feature	Genes		Exons	
Type	protein-coding	LncRNA	protein-coding	LncRNA
Total	19,384	101,285	476,299	611,102
Sense strand overlap	6309	7531	42,868	24,357
Antisense strand overlap	10,492	14,212	47,057	44,062
Post sense strand filtering	*	93,754	*	565,717
Post antisense strand filtering	*	91,215	*	537,373

Table 1. **Integrating GENCODE v32** (Frankish et al. 2019) **and LncExpDB v2** (Li et al. 2021)**.** GENCODE (dated 27.10.2021) contains 19,384 protein-coding genes (476,299 exons), and LncExpDB (27.10.2021) contains 101,285 lncRNA genes (611,102 exons). The table indicates the number of lncRNAs that were filtered based on sense strand and antisense overlap as described in the text. * denoted no filtering of protein-coding genes and exons.

Feature	Genes		Exons	
Type	protein-coding	LncRNA	protein-coding	LncRNA
Total	19,384	16,849	476,299	109,075
Sense strand overlap	619	516	3514	2106
Antisense strand overlap	3590	3791	12,941	8809

Table 2. **Distribution of protein-coding and lncRNA genes in GENCODE v32.**

Dataset	Source	Number of samples	Number of cells
pbmc_10k_v3 (10x Genomics)*	PBMCs	1	9432
GSE115469 (MacParland et al. 2018)	Liver	5	8444
GSE136103 (Ramachandran et al. 2019)	Liver	20	58358

Table 3. Datasets analyzed. 10x single cell RNA-seq datasets used to validate Singletrome annotation and create lncRNA cell type maps. * denotes 10k PBMCs from a Healthy Donor (v3 chemistry) Single Cell Gene Expression Dataset by Cell Ranger 3.0.0, 10x Genomics, (2018, November 19).

Material and methods

Genome indices. We downloaded the human reference genome index from 10x Genomics https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz, which includes genes from different biotypes (lncRNA, protein_coding, IG_V_pseudogene, IG_V_gene, IG_C_gene, IG_J_gene, TR_C_gene, TR_J_gene, TR_V_gene, TR_V_pseudogene, TR_D_gene, IG_C_pseudogene, TR_J_pseudogene, IG_J_pseudogene, IG_D_gene) as shown in Supplementary Table 6 along with the number of genes for each biotype. We termed this genome annotation as GENCODE (used by Cell Ranger) in the manuscript. For evaluating protein-coding and lncRNAs exonic overlap in the GENCODE annotation, we used the same strategy and script from 10x Genomics with protein_coding and lncRNA as the biotype patterns respectively. In brief, we obtained 19,384 protein-coding genes with GENCODE v32 filtering for 'protein_coding' as the 'gene_type' and 'transcript_type'. We additionally filtered transcripts with tags such as 'readthrough_transcript' and 'PAR'. We obtained 16,849 long noncoding RNAs filtering GENCODE v32 for 'lncRNA' as the 'gene_type' and 'transcript_type'. We additionally filtered transcripts with tags such as 'readthrough_transcript' and 'PAR'. For TLGA and ULGA genome indices, we downloaded the human lncRNA genome annotation file from ftp://download.big.ac.cn/lncexpdb/0-ReferenceGeneModel/1-GTFFiles/LncExpDB_OnlyLnc.tar.gz. We removed 8 genes (HSALNG0056858, HSALNG0059740, HSALNG0078365, HSALNG0092690, HSALNG009306, HSALNG0089130, HSALNG0089954 and HSALNG0095105) where we found invalid exons in the transcript or exons of transcripts were not stored in ascending order. To create the TLGA and ULGA genome indices, we included the protein-coding genes obtained from the GENCODE with the inhouse created genome annotation file (see section 'Expanding lncRNA annotations in single cell analysis'), and created the genome indices using the bash script available at 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build#hg19_3.0.0). For all the genome indices, the human reference sequence for GRCh38 was downloaded from http://ftp.ensembl.org/pub/release-98/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz. Genome indices were created using Cell Ranger version 3.1.0 due its compatibility with all the versions (3.1 to 6.0 as of the current work) of count pipelines and older v1 chemistry versions of Cell Ranger count (Supplementary Table 7). Using Cell Ranger version 3.1.0 mkref will help to analyze scRNA-seq data generated with the older v1 chemistry version.

Data. We analyzed three publicly available 10x scRNA-seq datasets consisting of 26 samples (Table 3) with the most widely used genome annotation for scRNA-seq analysis (GENCODE) and our custom genome annotations (TLGA and ULGA).

Gene expression. Cell Ranger count version 6.0.2 was used with the default parameters for all the genome versions to obtain gene expression count matrix.

lncRNA quality filter. To compute the gene body coverage for each dataset (PBMCs, liver set 1 and liver set 2) we utilized RSeQC (Wang, Wang, and Li 2012). The program was used to check if reads coverage was uniform and if there was any 5' or 3' end bias or if majority of the reads are mapped to one location (single bin) in the transcript. RSeQC scales all the transcripts to 100 bins and calculates the number of reads covering each bin position and provides the normalized coverage profile along the gene body. We modified the RSeQC geneBody_coverage.py script to obtain raw read counts (default is normalized read count to 1) for each bin. To assess the read distribution across the gene body and avoid transcript length bias, we subdivided lncRNAs and protein-coding transcripts based on transcript length. Gene and transcript length were calculated using R package GenomicFeatures version 1.46.1 (Supplementary Table 3). The input for the program is an indexed BAM file and gene model in BED format. Gene models were created for protein-coding genes from GENCODE and lncRNAs from Singletrome. We assessed read distribution across the transcript body to identify lncRNA genes where 1) mapped reads exhibit 5’ bias in 3’ sequenced scRNA-seq libraries and 2) the majority of reads were mapped to a single location in the transcript, as both situations could represent library artifacts or mapping anomalies (Ma and Kingsford 2019). LncRNA genes for which all transcripts met either criteria in a dataset were excluded from further analysis in that dataset. LncRNAs that passed these filtering steps were used for all the downstream analysis such as cell type clustering, cell type prediction, differential expression in healthy and cirrhotic liver and disease prediction.

Cell type clustering. We used Seurat version 4.0.6 for analyzing all the gene expression matrices for all the datasets (Table 3). We retained cell type labels from the original publications. We matched the barcodes from our mapping to the original publication barcodes to obtain the cell type labels. In all the analyses, the Singletrome count matrix was subsetted for protein-coding genes and lncRNAs to cluster cell types. Since we used the cell labels from the original publications (Table 3), we discarded all the other cells that were not labeled (assigned cell type identity) in the original publication.

Cell type markers identification. We used Seurat version 4.0.6 for the identification of cell type markers in all the datasets (Table 3). To identify cell type markers based on lncRNA and protein-coding genes, gene expression count matrices obtained from Singletrome mapping were split into protein-coding and lncRNA genes for each dataset. We used FindAllMarkers function from Seurat to find markers (differentially expressed genes) for each of the cell types in a dataset. We retained only those genes with a log-transformed fold change of at least 0.25 and expression in at least 25% of cells in the cluster under comparison.

Cell type prediction using machine learning. We trained a XGBoost classifier (version 1.6.2) on the expression data of protein-coding genes, lncRNAs, and the combination of both in Singletrome to predict cell types. Cell type labels were retained from the original publications (Table 3). We opted for XGBoost, as it is a preferential model for the imbalanced data and some cell types were under-represented in the datasets (Table 3 and Supplementary data 5-7). Expression data for each model (protein-coding, lncRNAs and Singletrome) was split into a training set (80%) and test set (20%). The model was trained using 80% of the data and evaluated using the remaining 20% of the data for each dataset (Table 3). To find the optimal parameters for the model, we used RandomizedSearchCV. The resultant optimal parameters for cell type classification were n_estimators : 25, max_depth : 25 and tree_method : 'hist'. Measurements of the model performance such as accuracy, recall, precision, f1, specificity, AUC are reported for each model for all the datasets (Supplementary data 5-7).

Differential expression analysis. We used Seurat version 4.0.6 to perform differential expression analysis between healthy and cirrhotic liver for liver set 2. The gene expression count matrix obtained from Singletrome mapping was split into protein-coding and lncRNA genes. Differential expression analysis was performed separately for Singletrome, protein-coding genes and lncRNA genes. We used FindMarkers function from Seurat to identify the differentially expressed genes between healthy and cirrhotic liver for each cell type. We filtered differentially expressed genes (protein-coding and lncRNAs) for padj-value less than 0.1 and log2FC more than 0.25 in either direction.

Disease (cirrhosis) prediction using machine learning. We trained XGBoost classifier (version 1.6.2) on the expression data of protein-coding genes, lncRNAs, and the combination of both in Singletrome to predict the condition (healthy or cirrhotic) of the cell in liver set 2.

Condition (healthy or cirrhotic) labels were retained from the original publication (liver set 2). RandomizedSearchCV technique was used to identify the optimum values of various parameters for the model. The optimum values obtained for various parameters were n_estimators: 400, max_depth: 25, subsample: 0.75, and tree_method:'hist'. Expression data for each model (protein-coding, lncRNAs and Singletrome) was split into a training set (80%) and test set (20%). The model was trained using 80% of the data and evaluated using the remaining 20% of the data. Measurements of the model performance such as accuracy, recall, precision, f1, specificity, AUC are reported for each model (Supplementary data 6).

Data availability. All the datasets (Table 3) used in this study are publicly available. The PBMCs dataset was obtained from the 10x Genomics platform "10k PBMCs from a Healthy Donor (v3 chemistry) Single Cell Gene Expression Dataset by Cell Ranger 3.0.0, 10x Genomics, (2018, November 19)". The previously-published datasets from the Gene Expression Omnibus (GEO) used in this study are GSE115469 and GSE136103.

Acknowledgments. The authors thank Kate Jeffrey for helpful discussion. A.C.M. was supported by the Chan Zuckerberg Initiative, Pew Biomedical Scholars Program, and NIH/NIDDK R01DK116999.

Author contributions. R.R and A.C.M. conceived and designed the study. Computational analysis was performed by R.R. I.A. and R.R. designed the cell type and disease prediction analysis and I.A implemented the Xgboost models. R.S and A.S assisted with the analysis of the differentially expressed lncRNAs in liver fibrosis. The manuscript was written by R.R. and A.C.M with input from all other authors.

Competing interests: A.C.M. receives research funding from Boehringer Ingelheim, Bristol-Myers Squibb, and Glaxo Smith Klein for other projects and is a consultant for Third Rock Ventures. R. R. is founder of deepnostiX in Germany and Pakistan.

References

Atanasovska, Biljana, Sander S. Rensen, Marijke R. van der Sijde, Glenn Marsman, Vinod Kumar, Iris Jonkers, Sebo Withoff, et al. 2017. “A Liver-Specific Long Noncoding RNA with a Role in Cell Viability Is Elevated in Human Nonalcoholic Steatohepatitis.” *Hepatology* 66 (3): 794–808.

Bell, Charles C., Paulo P. Amaral, Anton Kalsbeek, Graham W. Magor, Kevin R. Gillinder, Pierre Tangermann, Lorena di Lisio, et al. 2016. “The Evx1/Evx1as Gene Locus Regulates Anterior-Posterior Patterning during Gastrulation.” *Scientific Reports* 6 (May): 26657.

Bolha, Luka, Metka Ravnik-Glavač, and Damjan Glavač. 2017. “Long Noncoding RNAs as Biomarkers in Cancer.” *Disease Markers* 2017 (May): 7243968.

Cabili, M. N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn. 2011. “Integrative Annotation of Human Large Intergenic Noncoding RNAs Reveals Global Properties and Specific Subclasses.” *Genes & Development*. https://doi.org/10.1101/gad.17446611.

Daneshvar, Kaveh, M. Behfar Ardehali, Isaac A. Klein, Fu-Kai Hsieh, Arcadia J. Kratkiewicz, Amin Mahpour, Sabrina O. L. Cancelliere, et al. 2020. “lncRNA DIGIT and BRD3 Protein Form Phase-Separated Condensates to Regulate Endoderm Differentiation.” *Nature Cell Biology* 22 (10): 1211–22.

Delás, M. Joaquina, and Gregory J. Hannon. 2017. “lncRNAs in Development and Disease: From Functions to Mechanisms.” *Open Biology* 7 (7). https://doi.org/10.1098/rsob.170121.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. “The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression.” *Genome Research* 22 (9): 1775–89.

Ding, Jiarui, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, et al. 2020. “Systematic Comparison of Single-Cell and Single-Nucleus RNA-Sequencing Methods.” *Nature Biotechnology* 38 (6): 737–46.

Domínguez Conde, C., C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, et al. 2022. “Cross-Tissue Immune Cell Analysis Reveals Tissue-Specific Features in Humans.” *Science* 376 (6594): eabl5197.

Faghihi, Mohammad Ali, Ming Zhang, Jia Huang, Farzaneh Modarresi, Marcel P. Van der Brug, Michael A. Nalls, Mark R. Cookson, Georges St-Laurent 3rd, and Claes Wahlestedt. 2010. “Evidence for Natural Antisense Transcript-Mediated Inhibition of microRNA Function.” *Genome Biology* 11 (5): R56.

Fang, Shuangsang, Lili Zhang, Jincheng Guo, Yiwei Niu, Yang Wu, Hui Li, Lianhe Zhao, et al. 2018. “NONCODEV5: A Comprehensive Annotation Database for Long Non-Coding RNAs.” *Nucleic Acids Research* 46 (D1): D308–14.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.

Hernesniemi, Jussi A., Shadi Mahdiani, Juho A. Tynkkynen, Leo-Pekka Lyytikäinen, Pashupati P. Mishra, Terho Lehtimäki, Markku Eskola, Kjell Nikus, Kari Antila, and Niku Oksala. 2019. “Extensive Phenotype Data and Machine Learning in Prediction of Mortality in Acute Coronary Syndrome - the MADDEC Study.” *Annals of Medicine* 51 (2): 156–63.

Huang, Weizhen, Yunming Tian, Shaoting Dong, Yinlian Cha, Jun Li, Xiaohong Guo, and Xia Yuan. 2017. “The Long Non-Coding RNA SNHG3 Functions as a Competing Endogenous RNA to Promote Malignant Development of Colorectal Cancer.” *Oncology Reports* 38 (3): 1402–10.

Jiang, Lichun, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R.

Gingeras, and Brian Oliver. 2011. “Synthetic Spike-in Standards for RNA-Seq Experiments.” *Genome Research* 21 (9): 1543–51.

Kotzin, Jonathan J., Sean P. Spencer, Sam J. McCright, Dinesh B. Uthaya Kumar, Magalie A. Collet, Walter K. Mowel, Ellen N. Elliott, et al. 2016. “The Long Non-Coding RNA Morrbid Regulates Bim and Short-Lived Myeloid Cell Lifespan.” *Nature* 537 (7619): 239–43.

Liu, Siyuan John, Tomasz J. Nowakowski, Alex A. Pollen, Jan H. Lui, Max A. Horlbeck, Frank J. Attenello, Daniel He, et al. 2016. “Single-Cell Analysis of Long Non-Coding RNAs in the Developing Human Neocortex.” *Genome Biology* 17 (April): 67.

Li, Zhao, Lin Liu, Shuai Jiang, Qianpeng Li, Changrui Feng, Qiang Du, Dong Zou, Jingfa Xiao, Zhang Zhang, and Lina Ma. 2021. “LncExpDB: An Expression Database of Human Long Non-Coding RNAs.” *Nucleic Acids Research* 49 (D1): D962–68.

Ma, Cong, and Carl Kingsford. 2019. “Detecting, Categorizing, and Correcting Coverage Anomalies of RNA-Seq Quantification.” *Cell Systems* 9 (6): 589–99.e7.

MacParland, Sonya A., Jeff C. Liu, Xue-Zhong Ma, Brendan T. Innes, Agata M. Bartczak, Blair K. Gage, Justin Manuel, et al. 2018. “Single Cell RNA Sequencing of Human Liver Reveals Distinct Intrahepatic Macrophage Populations.” *Nature Communications* 9 (1): 1–21.

Mahpour, Amin, and Alan C. Mullen. 2021. “Our Emerging Understanding of the Roles of Long Non-Coding RNAs in Normal Liver Function, Disease, and Malignancy.” *JHEP Reports : Innovation in Hepatology* 3 (1): 100177.

Mourão, Kira, Nicholas J. Schurch, Radek Lucoszek, Kimon Froussios, Katarzyna MacKinnon, Céline Duc, Gordon Simpson, and Geoffrey J. Barton. 2019. “Detection and Mitigation of Spurious Antisense Expression with RoSA.” *F1000Research* 8 (June): 819.

Nath, Aritro, Eunice Y. T. Lau, Adam M. Lee, Paul Geeleher, William C. S. Cho, and R. Stephanie Huang. 2019. “Discovering Long Noncoding RNA Predictors of Anticancer Drug Sensitivity beyond Protein-Coding Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (44): 22020–29.

Nishio, Mizuho, Mitsuo Nishizawa, Osamu Sugiyama, Ryosuke Kojima, Masahiro Yakami, Tomohiro Kuroda, and Kaori Togashi. 2018. “Computer-Aided Diagnosis of Lung Nodule Using Gradient Tree Boosting and Bayesian Optimization.” *PloS One* 13 (4): e0195875.

Ogunleye, Adeola, and Qing-Guo Wang. 2020. “XGBoost Model for Chronic Kidney Disease Diagnosis.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* 17 (6): 2131–40.

Parkhomchuk, Dmitri, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch, Hans Lehrach, and Alexey Soldatov. 2009. “Transcriptome Analysis by Strand-Specific Sequencing of Complementary DNA.” *Nucleic Acids Research* 37 (18): e123.

Ramachandran, P., R. Dobie, J. R. Wilson-Kanamori, E. F. Dora, B. E. P. Henderson, N. T. Luu, J. R. Portman, et al. 2019. “Resolving the Fibrotic Niche of Human Liver Cirrhosis at Single-Cell Level.” *Nature* 575 (7783): 512–18.

Statello, Luisa, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. 2021. “Gene Regulation by Long Non-Coding RNAs and Its Biological Functions.” *Nature Reviews. Molecular Cell Biology* 22 (2): 96–118.

Suo, Chenqu, Emma Dann, Issac Goh, Laura Jardine, Vitalii Kleshchevnikov, Jong-Eun Park, Rachel A. Botting, et al. 2022. “Mapping the Developing Human Immune System across Organs.” *Science* 376 (6597): eabo0510.

Wang, Liguo, Shengqin Wang, and Wei Li. 2012. “RSeQC: Quality Control of RNA-Seq Experiments.” *Bioinformatics* 28 (16): 2184–85.

Wu, Xiong-Jian, Yuan Xie, Xiao-Xiang Gu, Hai-Yan Zhu, and Li-Xing Huang. 2022. “LncRNA XIST Promotes Mitochondrial Dysfunction of Hepatocytes to Aggravate Hepatic Fibrogenesis via miR-539-3p/ADAMTS5 Axis.” *Molecular and Cellular Biochemistry*, July. https://doi.org/10.1007/s11010-022-04506-0.

Xiao, Yongtao, Runping Liu, Xiaojiaoyang Li, Emily C. Gurley, Phillip B. Hylemon, Ying Lu, Huiping Zhou, and Wei Cai. 2019. “Long Noncoding RNA H19 Contributes to Cholangiocyte Proliferation and Cholestatic Liver Fibrosis in Biliary Atresia.” *Hepatology* 70 (5): 1658–73.

Yap, Kyoko L., Side Li, Ana M. Muñoz-Cabello, Selina Raguz, Lei Zeng, Shiraz Mujtaba, Jesús Gil, Martin J. Walsh, and Ming-Ming Zhou. 2010. “Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a.” *Molecular Cell* 38 (5): 662–74.

Zeng, Weihua, and Ali Mortazavi. 2012. “Technical Considerations for Functional Sequencing Assays.” *Nature Immunology* 13 (9): 802–7.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.

Main Figures

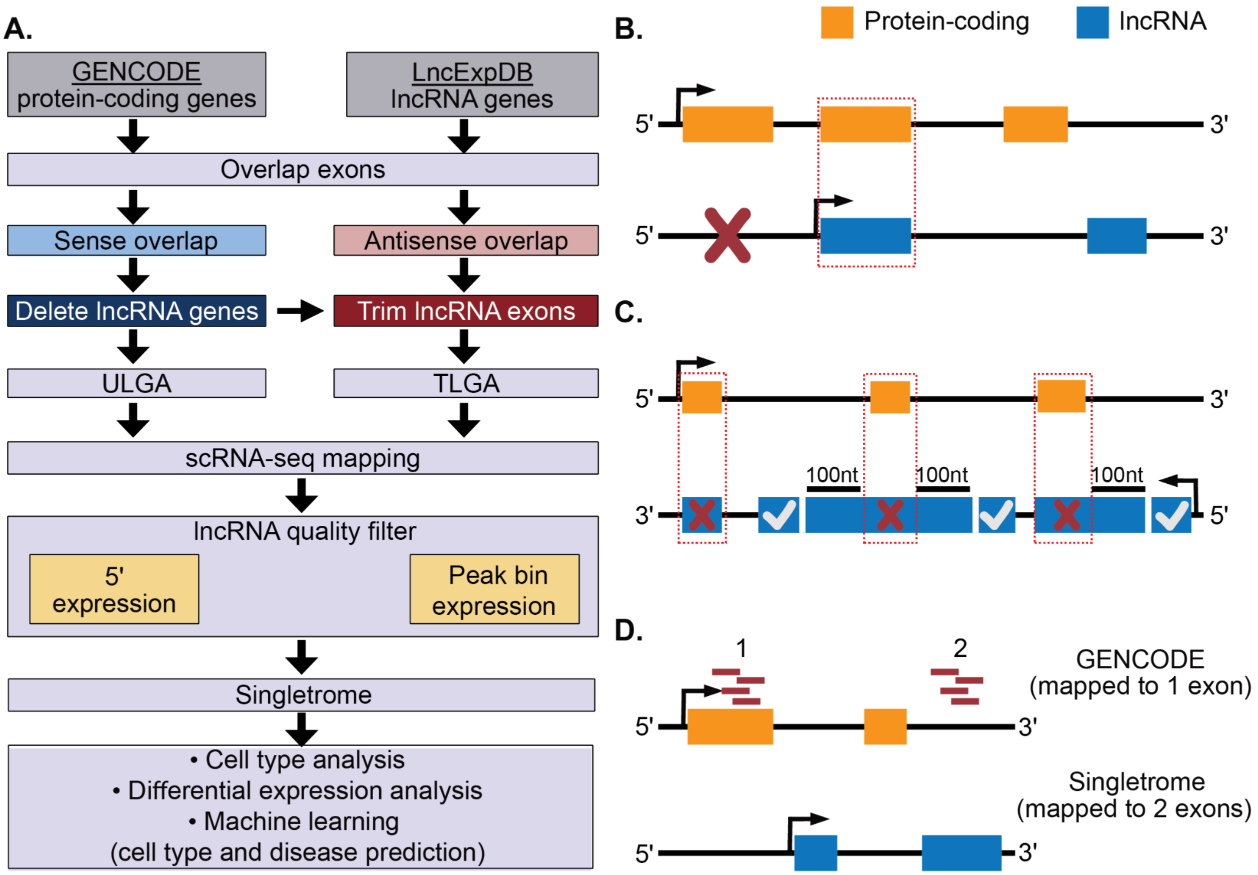


Figure 1. Enhancing the transcriptome with expanded lncRNA annotation for single cell analysis. (A) Development of Singletrome workflow. Exons of protein-coding genes from GENCODE v32 and lncRNA genes from LncExpDB v2 were integrated (Table 1). lncRNA genes overlapping on the sense strand with protein-coding genes were deleted to create the untrimmed lncRNA genome annotation (ULGA), and antisense strand overlapping lncRNA exons were trimmed to create the trimmed lncRNA genome annotation (TLGA) for scRNA-seq analysis. scRNA-seq data were mapped to both ULGA (to account for all lncRNA mapped reads) and TLGA (to define lncRNA expression based on reads with the highest confidence). Mapped lncRNAs were subjected to additional quality filters to remove transcripts that have reads mapped predominantly to the 5' end of a transcript or to a single, non 3' bin. The quality filtered lncRNAs (Singletrome) were used to perform cell type identification, differential expression analysis, and prediction of cell types and disease using machine learning. **(B)** Sense strand overlap. Cell Ranger discards reads mapped to overlapping exons on the same strand (red dotted box). To avoid miscounting reads to protein-coding genes by the inclusion of additional lncRNAs in the genome, lncRNA genes were discarded if they overlap in sense with protein-coding exons (red x), as it is more difficult to exclude the protein-coding potential of these lncRNAs. **(C)** Antisense strand overlap. Cell Ranger prioritizes alignments of sense over antisense reads. If spurious antisense reads are generated from transcripts of protein-coding genes, these could be incorrectly interpreted to indicate expression of an antisense overlapping lncRNA gene. To overcome this potential problem, we trimmed the overlapping region (red x) and an additional 100nt of lncRNA exons that were overlapping with protein-coding exons in the antisense direction. We retained the trimmed lncRNA exons if their length was at least 200nt (marked with white check). Gene and transcripts coordinates were updated accordingly. **(D)** Improved mapping specificity. Reads (red bars, 1) uniquely mapped to a single exon and were carried forward to UMI counting in GENCODE. In Singletrome, with the inclusion of 91,215 lncRNA genes (537,373 exons) these reads are now mapped to two exons (red bars 1 and 2). Removing these reads from UMI counting would improve read mapping specificity but reduce the total number of uniquely mapped reads.

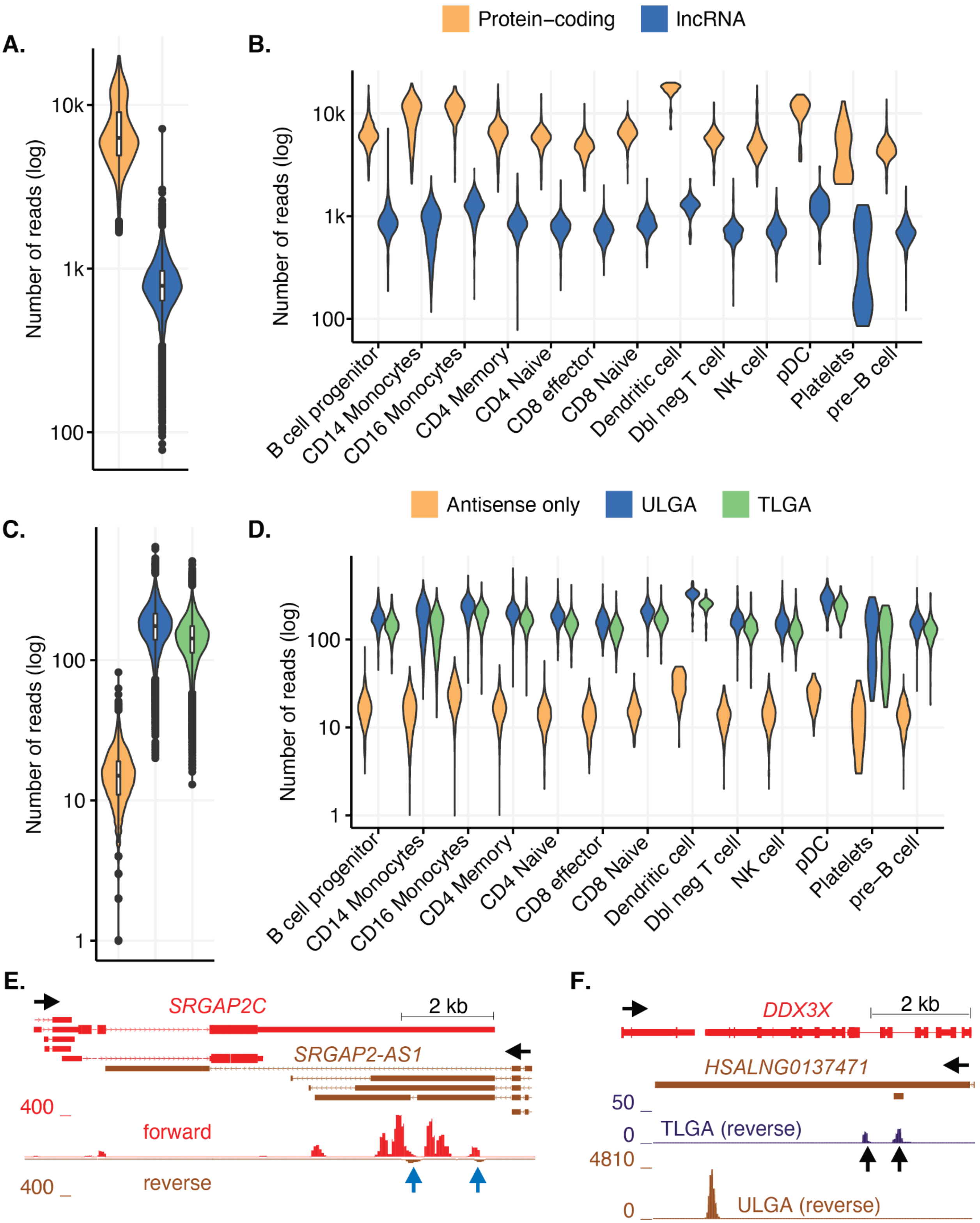

Protein−coding
lncRNA
A.
B.
Number of reads (log)
10k
1k
100
B cell progenitor
CD14 Monocytes
CD16 Monocytes
CD4 Memory
CD4 Naive
CD8 effector
CD8 Naive
Dendritic cell
Dbl neg T cell
NK cell
pDC
Platelets
pre−B cell
Antisense only
ULGA
TLGA
C.
D.
100
10
1
E.
SRGAP2C
2 kb
SRGAP2-AS1
400
forward
reverse
400
F.
DDX3X
2 kb
HSALNG0137471
50
0
TLGA (reverse)
4810
0
ULGA (reverse)

Figure 2. Distribution of transcripts in bulk and by cell type in PBMCs. (A) lncRNAs (blue) are expressed at lower levels than protein-coding genes (orange) in PBMCs. Reads aligned to lncRNAs and protein-coding genes are shown in log scale (y-axis). **(B)** lncRNAs are expressed at lower levels than protein-coding genes in all PBMC cell types. **(C)** Expression of lncRNAs that overlap protein-coding genes in the antisense direction. Expression of lncRNAs in non-overlapping regions [TLGA (green)], expression of lncRNAs in overlapping and non-overlapping regions [ULGA (blue)], and lncRNA exons that are expressed only in regions antisense to protein-coding exons [antisense only (orange)] are displayed (y-axis). The Y-axis shows the number of reads in log scale. TLGA (green) identifies lncRNAs with the highest confidence in expression but reduces the reads associated with lncRNAs as compared to ULGA (blue). **(D)** Data for each cell type are shown as described in (C). **(E)** Example where using ULGA reads could incorrectly suggest expression of an antisense lncRNA gene. *SRGAP2-AS1* (brown) is expressed antisense to *SRGAP2C* (red). Reads mapped to *SRGAP2C* (forward) are shown in red, while ULGA reads mapped to *SRGAP2-AS1* (reverse) are shown in brown. In this example, the reads mapped to *SRGAP2-AS1* are contained in exons antisense to an exon of *SRGAP2C,* where there are many more reads supporting the mRNA antisense to the lncRNA (blue arrows). This lncRNA gene is discarded because there are insufficient reads to support expression of *SRGAP2-AS1* in regions that do not overlap with exons of *SRGAP2 (SRGAP2-AS1* has no reads mapped in *TLGA)*. The genomic scale is indicated on the upper right, and the direction of transcription is indicated by horizontal black arrows. **(F)** Example where TLGA identifies reads mapped to exons of *HSALNG0137471* that do not overlap (in antisense) to exons of *DDX3X* (black vertical arrows), but does not capture the majority of reads mapped towards the 3’ end of *HSALNG0137471*. LncRNA *HSALNG0137471* (brown) is antisense to *DDX3X* (red). TLGA (reverse) only displays reads mapped to *HSALNG0137471* in exons that do not overlap (in antisense) to exons of *DDX3X*. ULGA (reverse) shows all reads mapped to *HSALNG0137471*. In this case, all reads mapped to *HSALNG0137471* [ULGA (reverse)] are used for down-stream analysis after the lncRNA is defined as expressed based on TLGA reads.

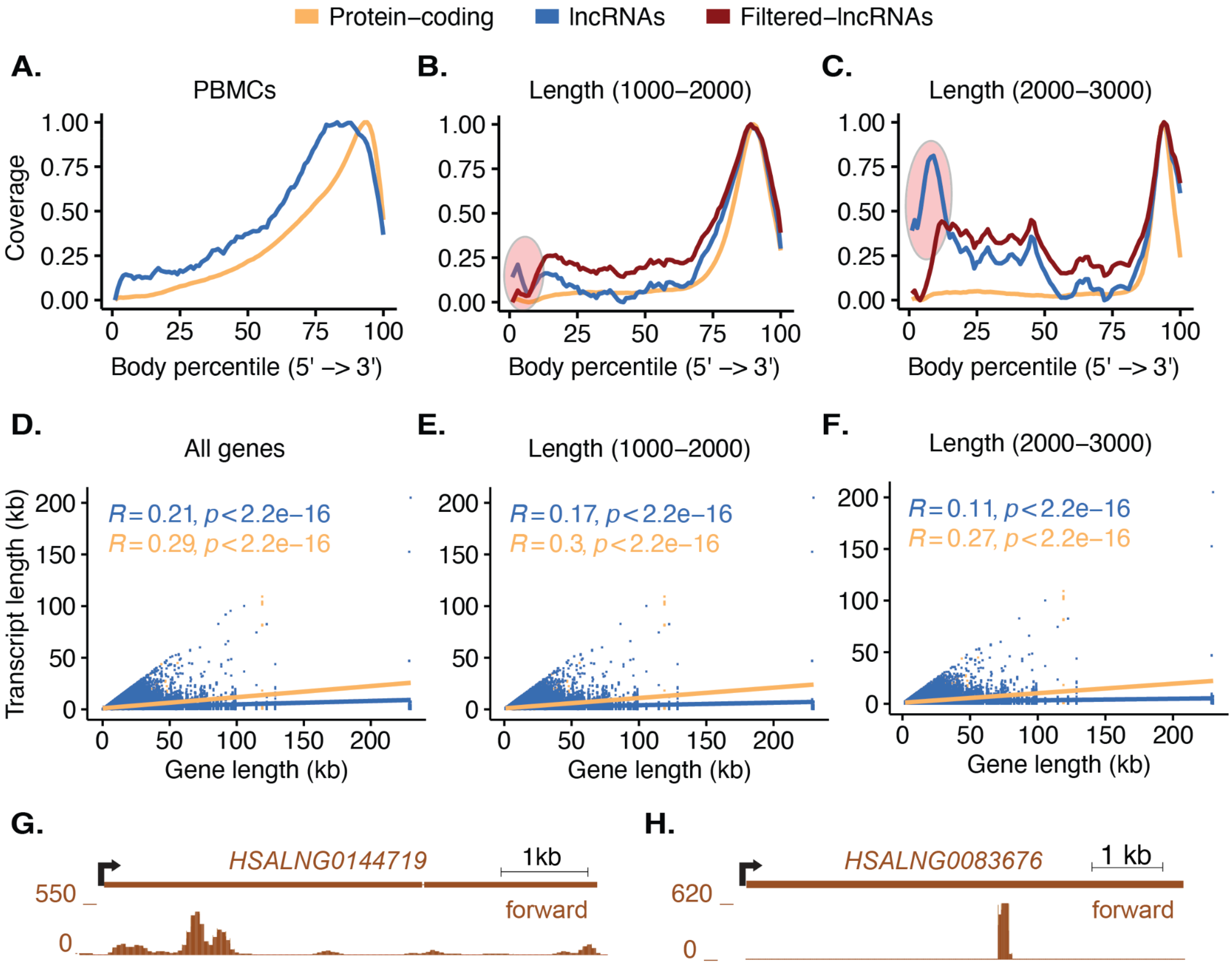

Protein−coding
lncRNAs
Filtered−lncRNAs
A.
PBMCs
B.
Length (1000−2000)
C.
Length (2000−3000)
Coverage
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Body percentile (5' −> 3')
D.
All genes
E.
Length (1000−2000)
F.
Length (2000−3000)
Transcript length (kb)
200
150
100
50
0
$R = 0.21$, $p < 2.2e-16$
$R = 0.29$, $p < 2.2e-16$
$R = 0.17$, $p < 2.2e-16$
$R = 0.3$, $p < 2.2e-16$
$R = 0.11$, $p < 2.2e-16$
$R = 0.27$, $p < 2.2e-16$
Gene length (kb)
G.
HSALNG0144719
1kb
550 _
0 _
forward
H.
HSALNG0083676
1 kb
620 _
0 _
forward

Figure 3. Distribution and quality control of lncRNA mapping in PBMCs. **(A)** Distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue). The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates transcript coverage (0-1). The overall read distribution for lncRNA genes is similar to protein-coding genes when all transcripts are considered. **(B)** Distribution of reads mapped across transcripts from 1000-2000 nt in length. Red circle indicates an enrichment of reads in the first 10 bins of lncRNA transcripts. The transcripts responsible for this peak were identified and filtered. Filtered-lncRNAs (red line) shows the distribution of the mapped reads after removing lncRNAs that were flagged for low quality (material and methods). **(C)** Distribution of reads mapped across transcripts from 2000-3000 nt in length. Red circle indicates an enrichment of reads in the first 10 bins of lncRNA transcripts. The transcripts responsible for this peak were identified and filtered. Filtered-lncRNAs (red line) shows the distribution of the mapped reads after removing lncRNAs that were flagged for low quality (material and methods). **(D)** The correlation between transcript length (y-axis) and gene length (x-axis) is weaker for lncRNA genes (blue) than protein-coding genes (orange). Gene length (x-axis) is plotted versus transcript length (y-axis) for all lncRNAs (blue dots). The blue line indicates R, and the R value is displayed in blue. The orange line represents R for protein-coding genes, and the R value is displayed in orange. **(E)** The correlation between transcript length (y-axis) and gene length (x-axis) is plotted as in (D) for all protein-coding genes and lncRNA genes with at least one transcript with length between 1000 and 2000 nt. **(F)** The correlation between transcript length (y-axis) and gene length (x-axis) is plotted as in (D) for all protein-coding genes and lncRNA genes with at least one transcript with length between 2000 and 3000 nucleotides. **(G)** Example where reads mapped to lncRNA gene *HSALNG0144719* show that the majority of reads are from the 5' end of the transcript and do not follow the expected distribution towards the 3' end of the transcript. This lncRNA gene is discarded. The genomic scale is indicated on the upper right. The start site and direction of transcription are indicated by a black arrow. **(H)** Example where lncRNA gene *HSALNG0083676* has the majority of reads mapped to a single location in the transcript and this location is not at the 3' end (last 10 bins). This lncRNA gene is discarded because this could be a library artifact or mapping anomaly.

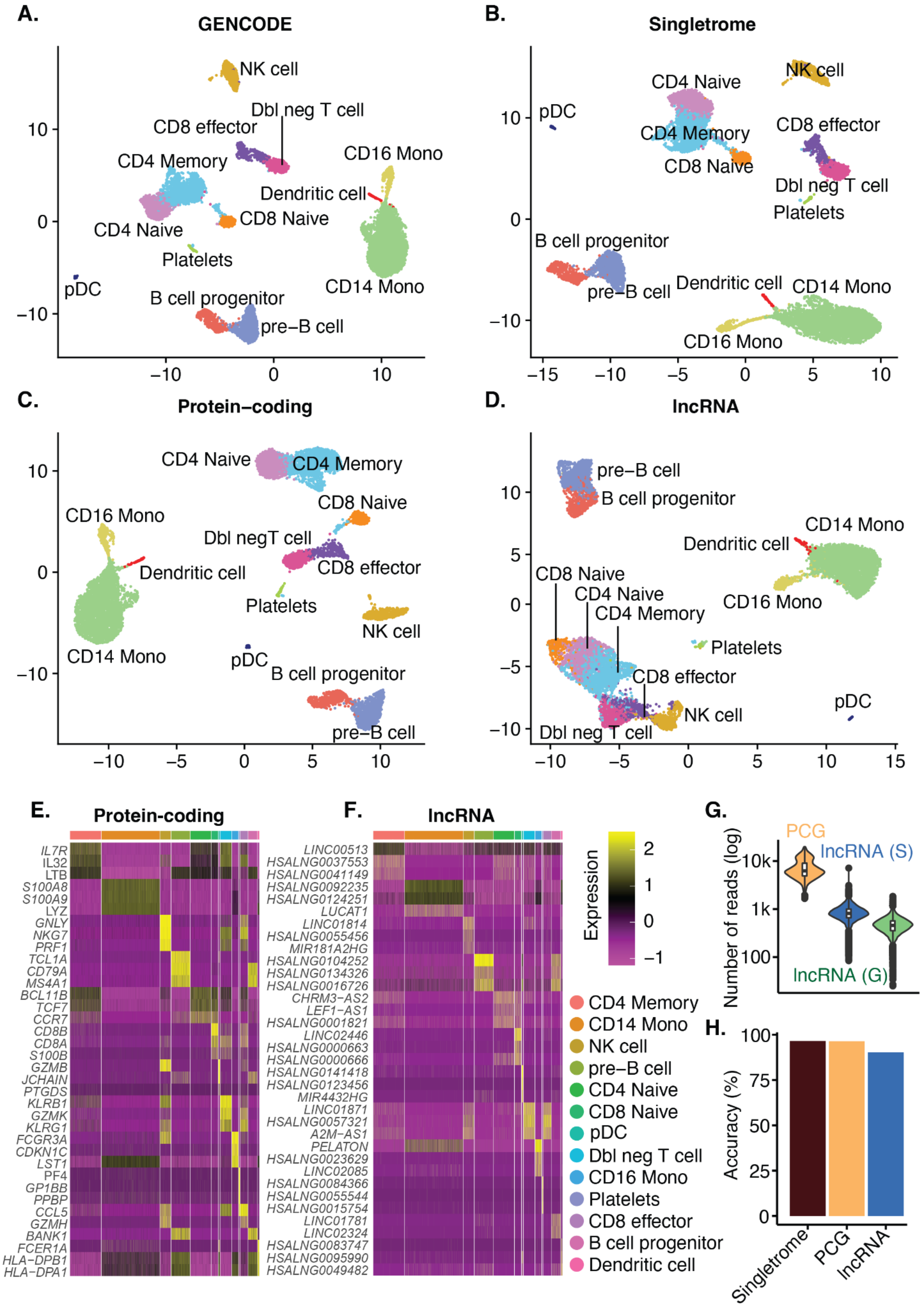
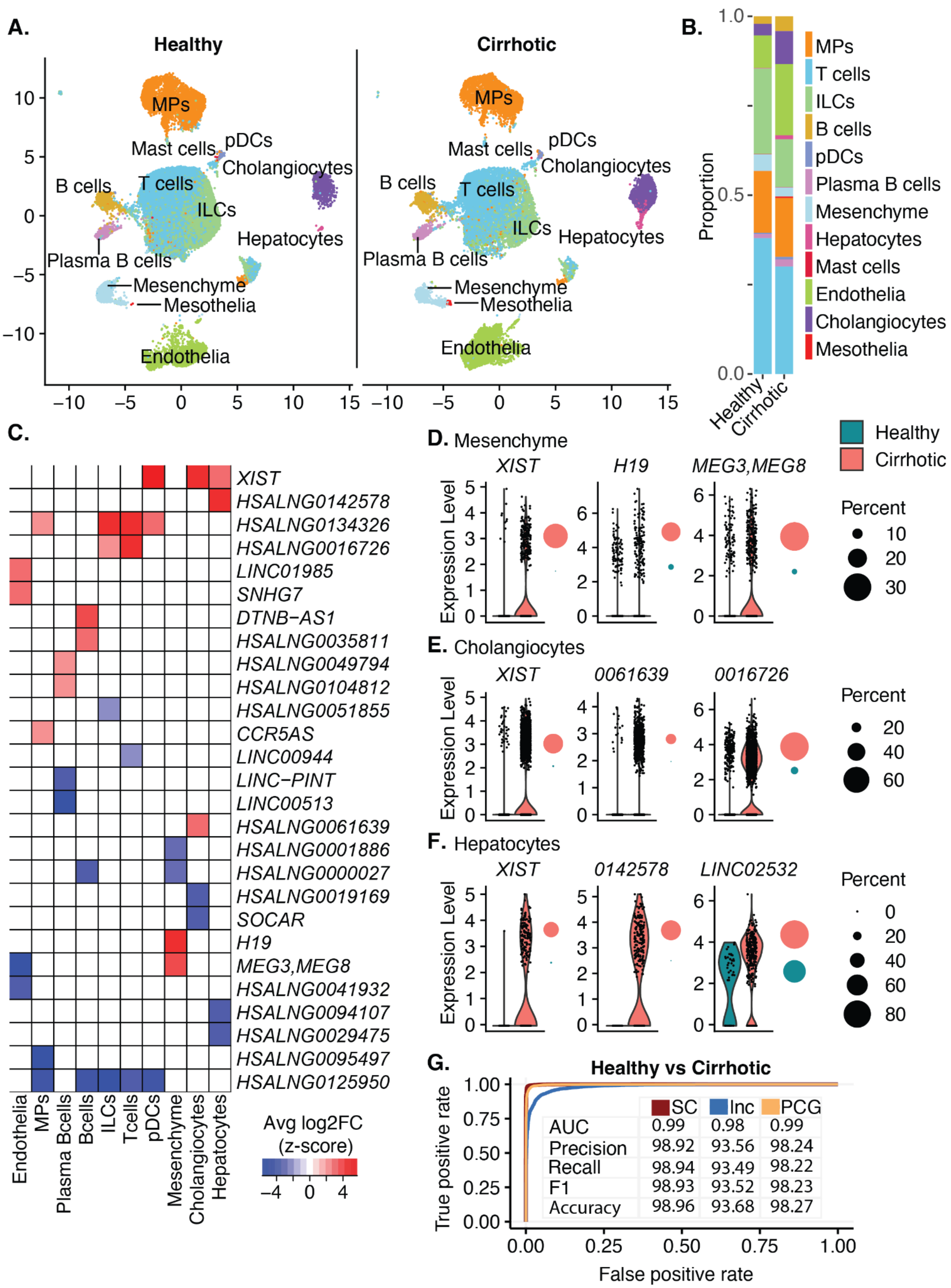

A.
GENCODE
NK cell
Dbl neg T cell
CD8 effector
CD16 Mono
CD4 Memory
Dendritic cell
CD8 Naive
CD4 Naive
Platelets
pDC
B cell progenitor
CD14 Mono
pre−B cell
10
0
−10
−10
0
10
B.
Singletrome
CD4 Naive
NK cell
pDC
CD4 Memory
CD8 effector
CD8 Naive
Dbl neg T cell
Platelets
B cell progenitor
pre−B cell
Dendritic cell
CD14 Mono
CD16 Mono
10
0
−10
−15
−10
−5
0
5
10
C.
Protein−coding
CD4 Naive
CD4 Memory
CD8 Naive
CD16 Mono
Dbl negT cell
Dendritic cell
CD8 effector
Platelets
NK cell
CD14 Mono
pDC
B cell progenitor
pre−B cell
10
0
−10
−10
−5
0
5
10
D.
lncRNA
pre−B cell
B cell progenitor
CD14 Mono
Dendritic cell
CD8 Naive
CD4 Naive
CD4 Memory
CD16 Mono
Platelets
CD8 effector
NK cell
pDC
Dbl neg T cell
10
5
0
−5
−10
−10
−5
0
5
10
15
E.
Protein-coding
IL7R
IL32
LTB
S100A8
S100A9
LYZ
GNLY
NKG7
PRF1
TCL1A
CD79A
MS4A1
BCL11B
TCF7
CCR7
CD8B
CD8A
S100B
GZMB
JCHAIN
PTGDS
KLRB1
GZMK
KLRG1
FCGR3A
CDKN1C
LST1
PF4
GP1BB
PPBP
CCL5
GZMH
BANK1
FCER1A
HLA−DPB1
HLA−DPA1
F.
lncRNA
LINC00513
HSALNG0037553
HSALNG0041149
HSALNG0092235
HSALNG0124251
LUCAT1
LINC01814
HSALNG0055456
MIR181A2HG
HSALNG0104252
HSALNG0134326
HSALNG0016726
CHRM3−AS2
LEF1−AS1
HSALNG0001821
LINC02446
HSALNG0000663
HSALNG0000666
HSALNG0141418
HSALNG0123456
MIR4432HG
LINC01871
HSALNG0057321
A2M−AS1
PELATON
HSALNG0023629
LINC02085
HSALNG0084366
HSALNG0055544
HSALNG0015754
LINC01781
LINC02324
HSALNG0083747
HSALNG0095990
HSALNG0049482
Expression
2
1
0
−1
● CD4 Memory
● CD14 Mono
● NK cell
● pre−B cell
● CD4 Naive
● CD8 Naive
● pDC
● Dbl neg T cell
● CD16 Mono
● Platelets
● CD8 effector
● B cell progenitor
● Dendritic cell
G.
PCG
lncRNA (S)
lncRNA (G)
Number of reads (log)
10k
1k
100
H.
Accuracy (%)
100
75
50
25
0
Singletrome
PCG
lncRNA

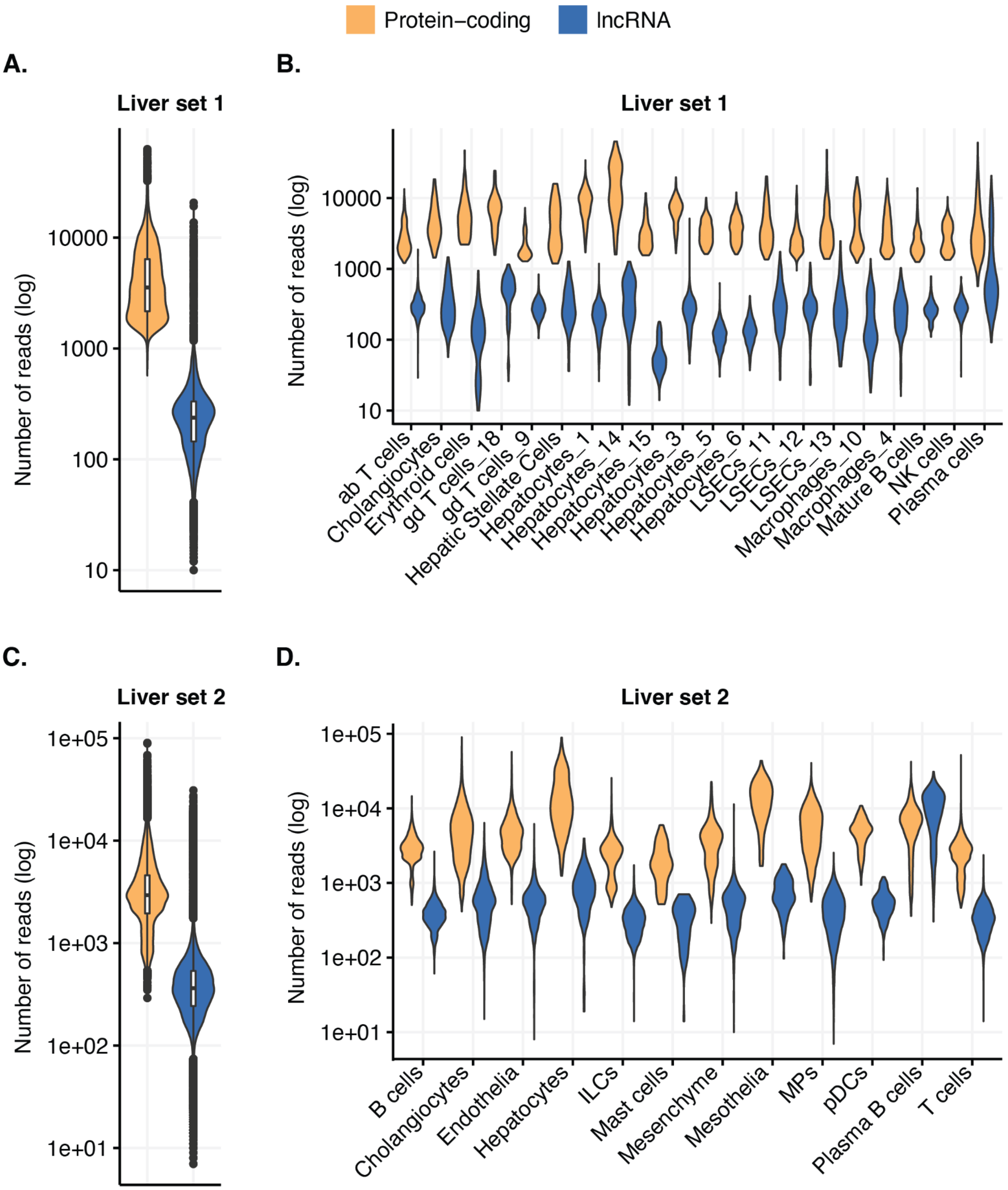
Figure 4. lncRNAs alone predict most clusters and cell types in single cell data. scRNA-seq data of PBMCs were mapped using annotation from **(A)** GENCODE, **(B)** Singletrome, **(C)** only protein-coding genes in Singletrome, and **(D)** only lncRNAs in Singletrome. The labels for each cell were retained from the original publications. For this analysis, Singletrome only contains lncRNAs that meet all described filters for PBMCs. **(E)** The heatmap displays the top differentially expressed protein-coding genes (y-axis) for each cell type in PBMCs. Cell types are indicated by color at the bar above the heatmap, and the key is displayed to the right. Expression level is indicated by Z-score. **(F)** The heatmap displays the top differentially expressed lncRNA genes for each cell type using the same gene expression scale as (E). Monocyte is abbreviated as Mono and double negative T cell is abbreviated as Dbl neg T cell. **(G)** The total number of mapped reads per cell (y-axis, log scale) is quantified for PCG (protein-coding genes) (orange), lncRNA (S) (lncRNA genes from Singletrome) (blue), and lncRNA (G) (lncRNA genes from GENCODE) (green) in PMBCs. **(H)** Bars showing accuracy in percentage (y-axis) for PBMC cell type prediction based on Singletrome (dark-red), PCG (protein-coding genes) from Singletrome (orange), and lncRNAs from Singletrome (blue). Receiver-operating characteristic (ROC) curves for each cell type are shown in Supplementary Fig 16.



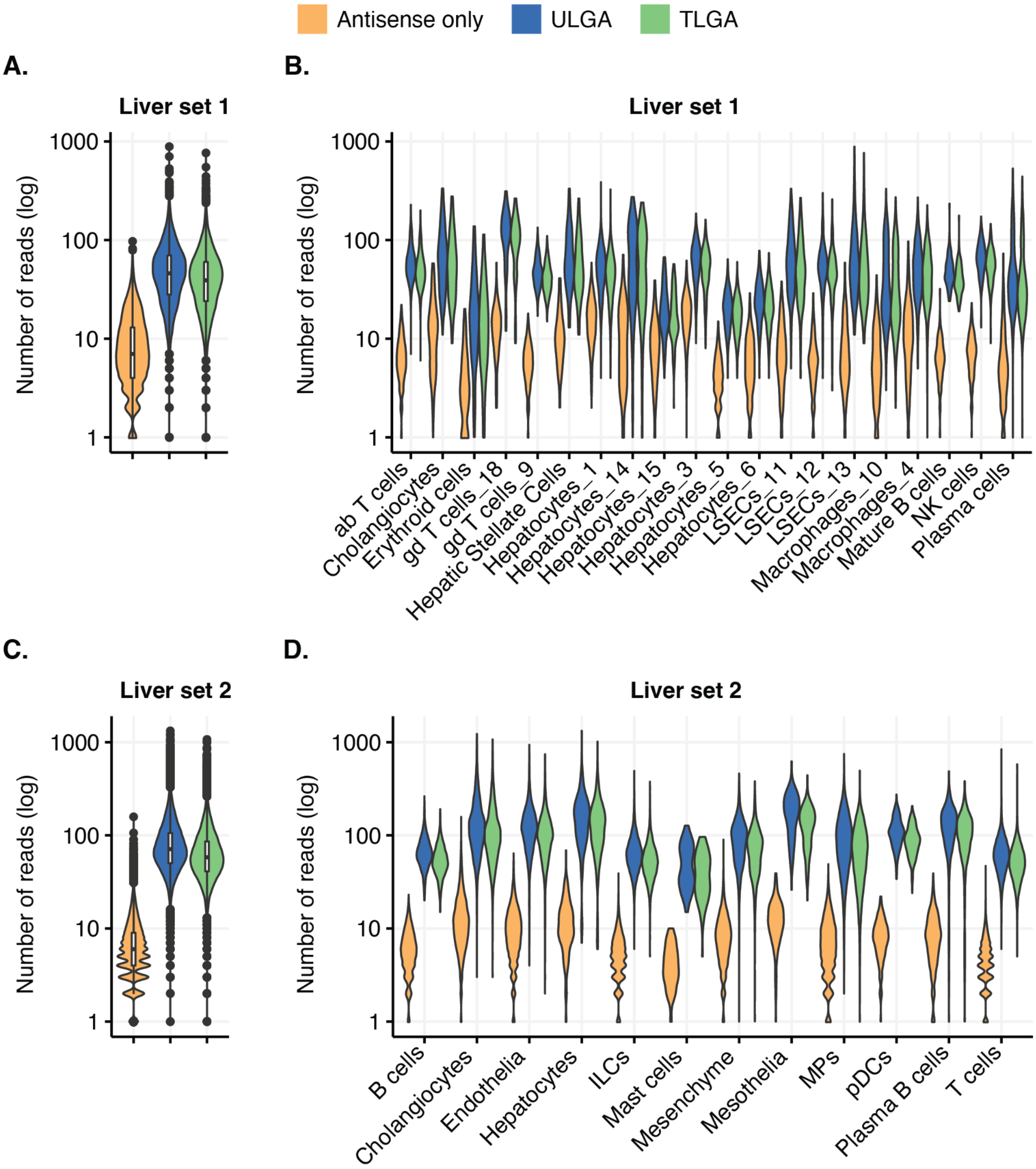
	SC	lnc	PCG
AUC	0.99	0.98	0.99
Precision	98.92	93.56	98.24
Recall	98.94	93.49	98.22
F1	98.93	93.52	98.23
Accuracy	98.96	93.68	98.27

Figure 5. lncRNA expression predicts disease pathology. (A) scRNA-seq data of liver set 2 (GSE136103(Ramachandran et al. 2019)) were mapped using annotations from Singletrome. The labels for each cell were retained from the original publication. Cells were clustered based on lncRNAs from Singletrome and annotated by condition healthy (left) and cirrhotic liver (right). For this analysis, only lncRNAs that meet all the described filters in the section ‘Quality control of lncRNA mapping’ are considered. **(B)** Proportion of cells in each cell type, healthy (left) and cirrhotic liver (right). **(C)** The heatmap displays the top differentially expressed (two up- and two down-regulated) lncRNA genes (y-axis) between healthy and cirrhotic liver for each cell type. Average log2-fold is indicated by Z-score for each cell type. **(D)** Differentially regulated lncRNA expression in mesenchymal cells, **(E)** cholangiocytes and **(F)** hepatocytes. Y-axis shows the expression of the differentially regulated lncRNA gene in cells of healthy and cirrhotic liver. Circles on the right show the percentage of cells expressing the lncRNA gene in healthy (green) and cirrhotic liver (red). **(G)** Receiver-operating characteristic (ROC) curve showing true and false positive rates for healthy and cirrhotic disease prediction based on the expression of SC (all genes in Singletrome) (red), lnc (lncRNA genes) (blue) and PCG (protein-coding genes) (orange). The table shows the AUC, precision (%), recall (%), F1 (%), and accuracy (%) for healthy and cirrhotic disease prediction based on the expression of SC (all genes in Singletrome) (red), lnc (lncRNA genes) (blue) and PCG (protein-coding genes) (orange).

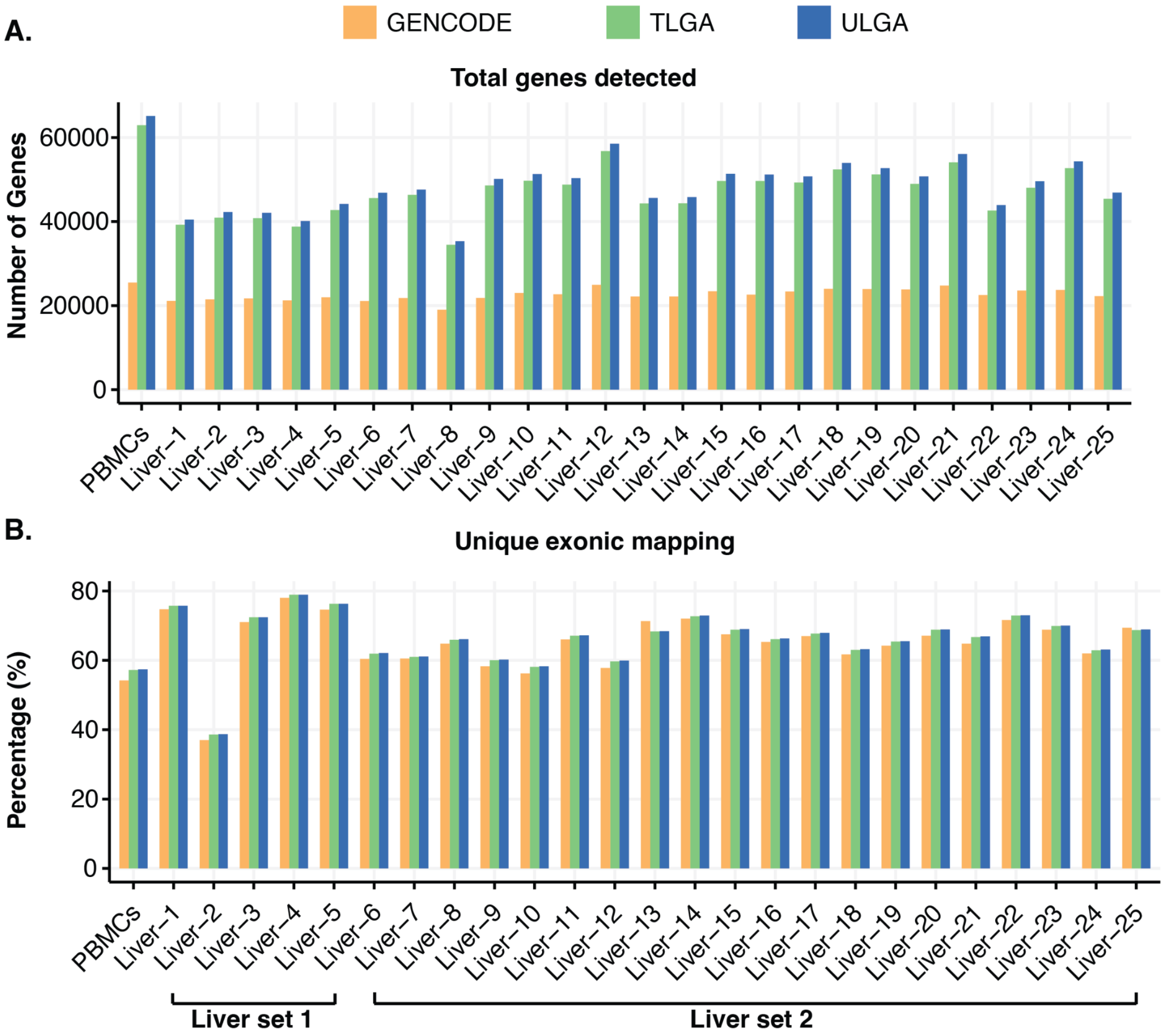
Supplementary Figures



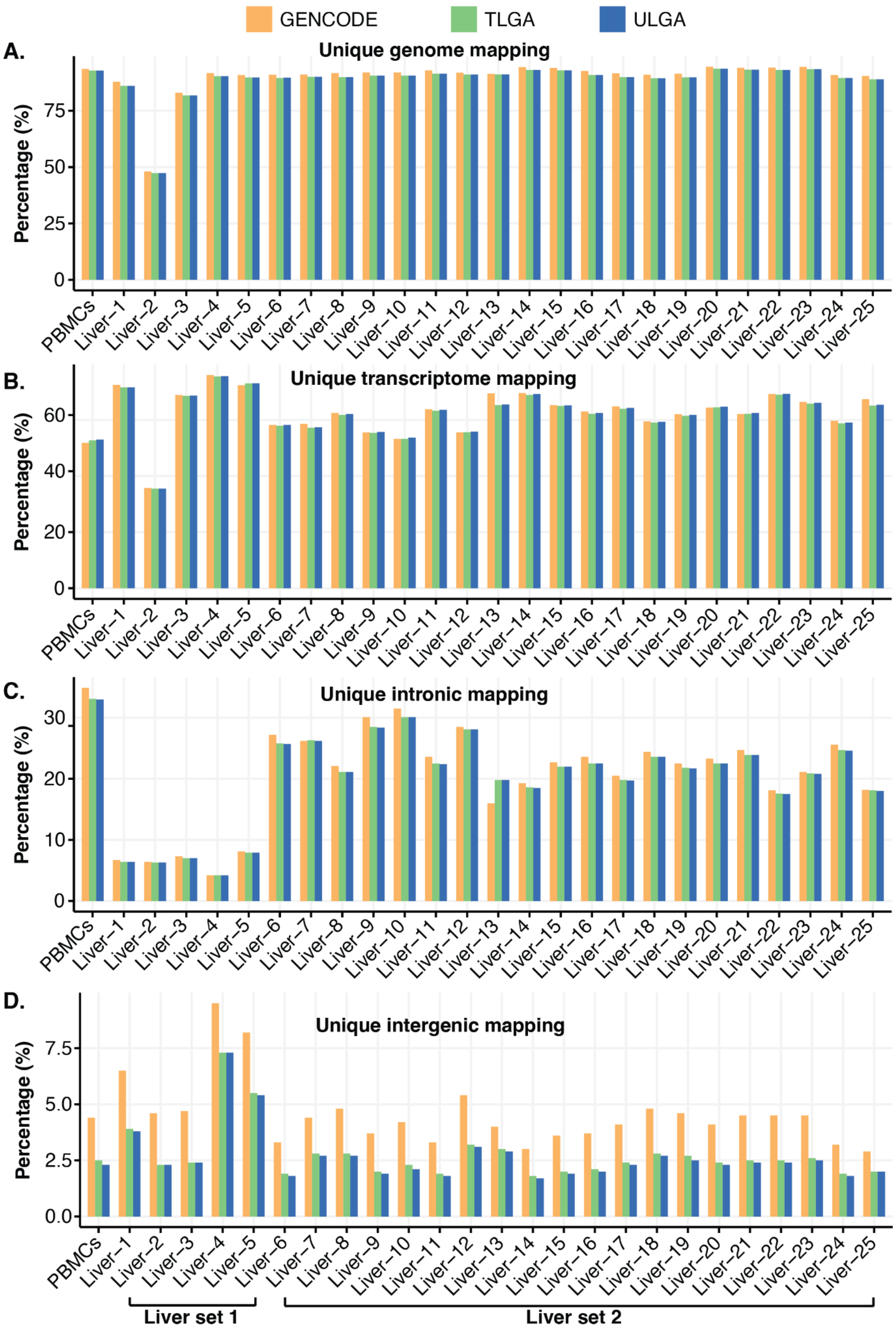
Supplementary Figure 1. Distribution of transcripts in bulk and by cell type in liver. LncRNAs (blue) are expressed at lower levels than protein-coding genes (orange) in liver set 1 **(A),** all cell types of liver set 1 **(B),** liver set 2 **(C),** and all cell types of liver set 2 **(D)**. Reads aligned to lncRNAs and protein-coding genes are shown in log scale (y-axis). Cell types (x-axis) are determined by the original description from liver set 1 (GSE115469 (MacParland et al. 2018)) and liver set 2 (GSE136103 (Ramachandran et al. 2019)).


Antisense only
ULGA
TLGA
A.
Liver set 1
Number of reads (log)
1000
100
10
1
B.
Liver set 1
Number of reads (log)
1000
100
10
1
ab T cells
Cholangiocytes
Erythroid cells
gd T cells_18
gd T cells_9
Hepatic Stellate Cells
Hepatocytes_1
Hepatocytes_14
Hepatocytes_15
Hepatocytes_3
Hepatocytes_5
Hepatocytes_6
LSECs_11
LSECs_12
LSECs_13
Macrophages_10
Macrophages_4
Mature B cells
NK cells
Plasma cells
C.
Liver set 2
Number of reads (log)
1000
100
10
1
D.
Liver set 2
Number of reads (log)
1000
100
10
1
B cells
Cholangiocytes
Endothelia
Hepatocytes
ILCs
Mast cells
Mesenchyme
Mesothelia
MPs
pDCs
Plasma B cells
T cells

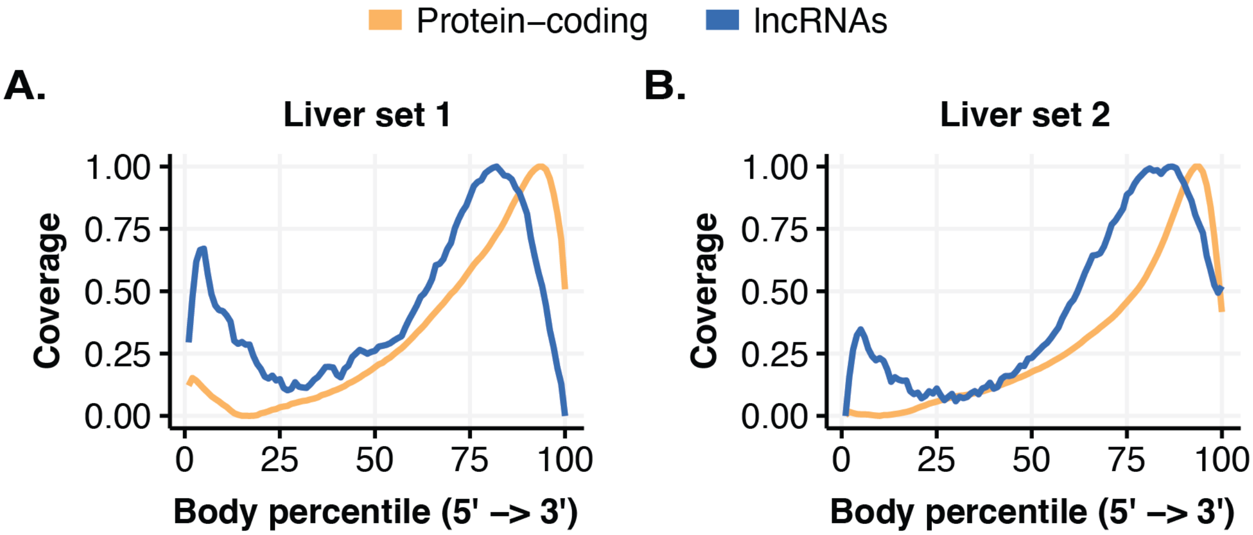
Supplementary Figure 2. Loss of lncRNA expression due to trimming in bulk and by cell type in liver. Expression of lncRNAs that overlap protein-coding genes in the antisense direction. Expression of lncRNAs in non-overlapping regions [TLGA (green)], expression of lncRNAs in overlapping and non-overlapping regions [ULGA (blue)], and lncRNA exons that are expressed only in regions antisense to protein-coding exons [antisense only (orange)] are displayed (y-axis). The Y-axis shows the number of reads in log scale. TLGA (green) identifies lncRNAs with the highest confidence in expression but reduces the reads associated with lncRNAs as compared to ULGA (blue) in liver set 1 **(A),** all cell types of liver set 1 **(B),** liver set 2 **(C)** and all cell types of liver set 2 **(D)**. Cell types (x-axis) are determined by the original description from liver set 1 (GSE115469 (MacParland et al. 2018)) and liver set 2 (GSE136103 (Ramachandran et al. 2019)).


A.
GENCODE
TLGA
ULGA
Total genes detected
Number of Genes
0
20000
40000
60000
PBMCs
Liver−1
Liver−2
Liver−3
Liver−4
Liver−5
Liver−6
Liver−7
Liver−8
Liver−9
Liver−10
Liver−11
Liver−12
Liver−13
Liver−14
Liver−15
Liver−16
Liver−17
Liver−18
Liver−19
Liver−20
Liver−21
Liver−22
Liver−23
Liver−24
Liver−25
B.
Unique exonic mapping
Percentage (%)
0
20
40
60
80
Liver set 1
Liver set 2

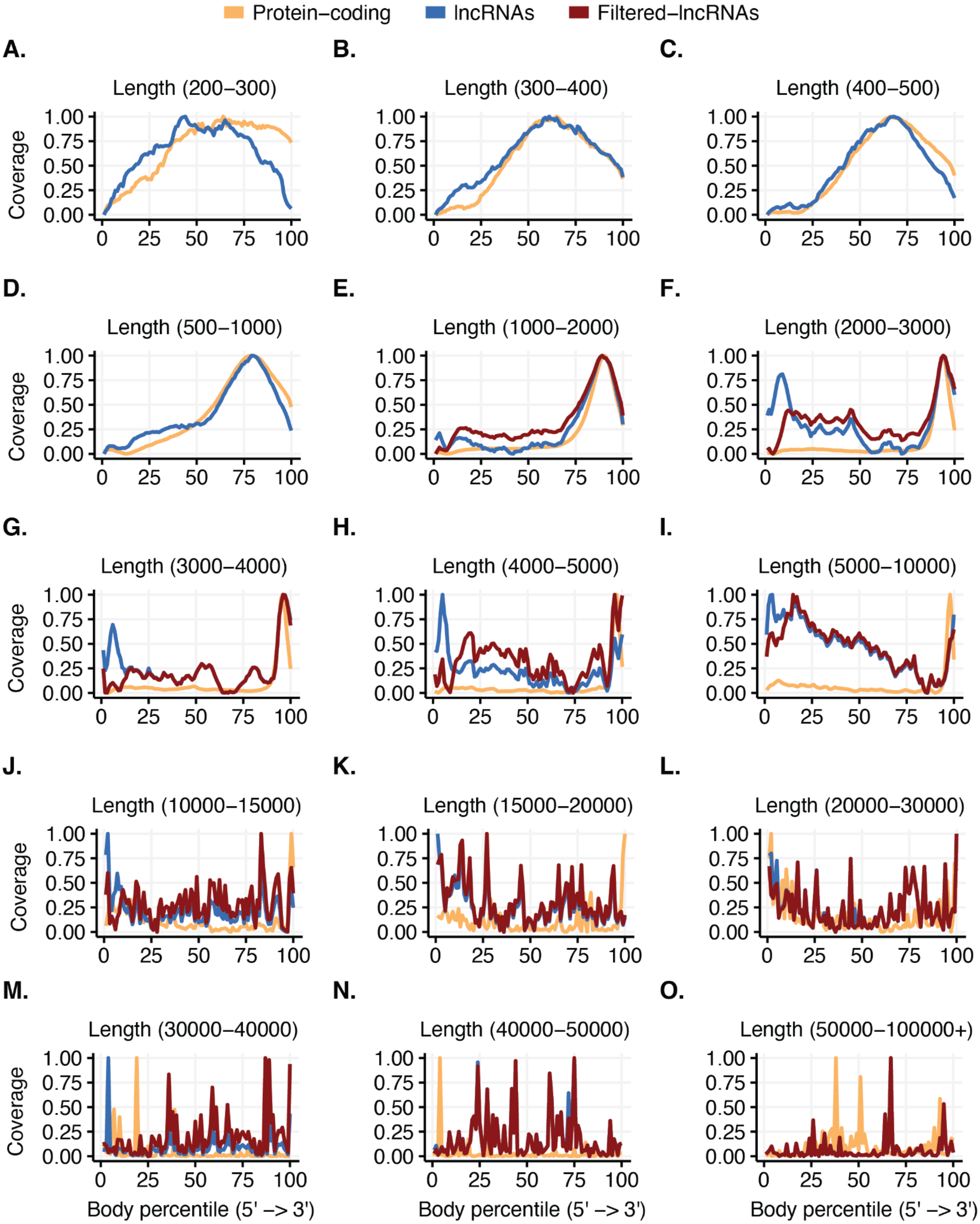
Supplementary Figure 3. Comparison of total genes detected and unique exonic mapping with GENCODE, TLGA, and ULGA annotations. (A) The total number of genes mapped (y-axis) increases in TLGA and ULGA compared to GENCODE. Analysis was performed for PBMCs (10x genomics) and liver set 1 (liver 1-5) (MacParland et al. 2018) and liver set 2 (liver 6-25) (Ramachandran et al. 2019). **(B)** The percentage of reads mapped uniquely to exons in TLGA and ULGA compared to GENCODE in PBMCs, samples from liver set 1, and liver set 2.


GENCODE
TLGA
ULGA
A.
Unique genome mapping
Percentage (%)
0
25
50
75
B.
Unique transcriptome mapping
0
20
40
60
C.
Unique intronic mapping
0
10
20
30
D.
Unique intergenic mapping
0.0
2.5
5.0
7.5
PBMCs
Liver−1
Liver−2
Liver−3
Liver−4
Liver−5
Liver−6
Liver−7
Liver−8
Liver−9
Liver−10
Liver−11
Liver−12
Liver−13
Liver−14
Liver−15
Liver−16
Liver−17
Liver−18
Liver−19
Liver−20
Liver−21
Liver−22
Liver−23
Liver−24
Liver−25
Liver set 1
Liver set 2

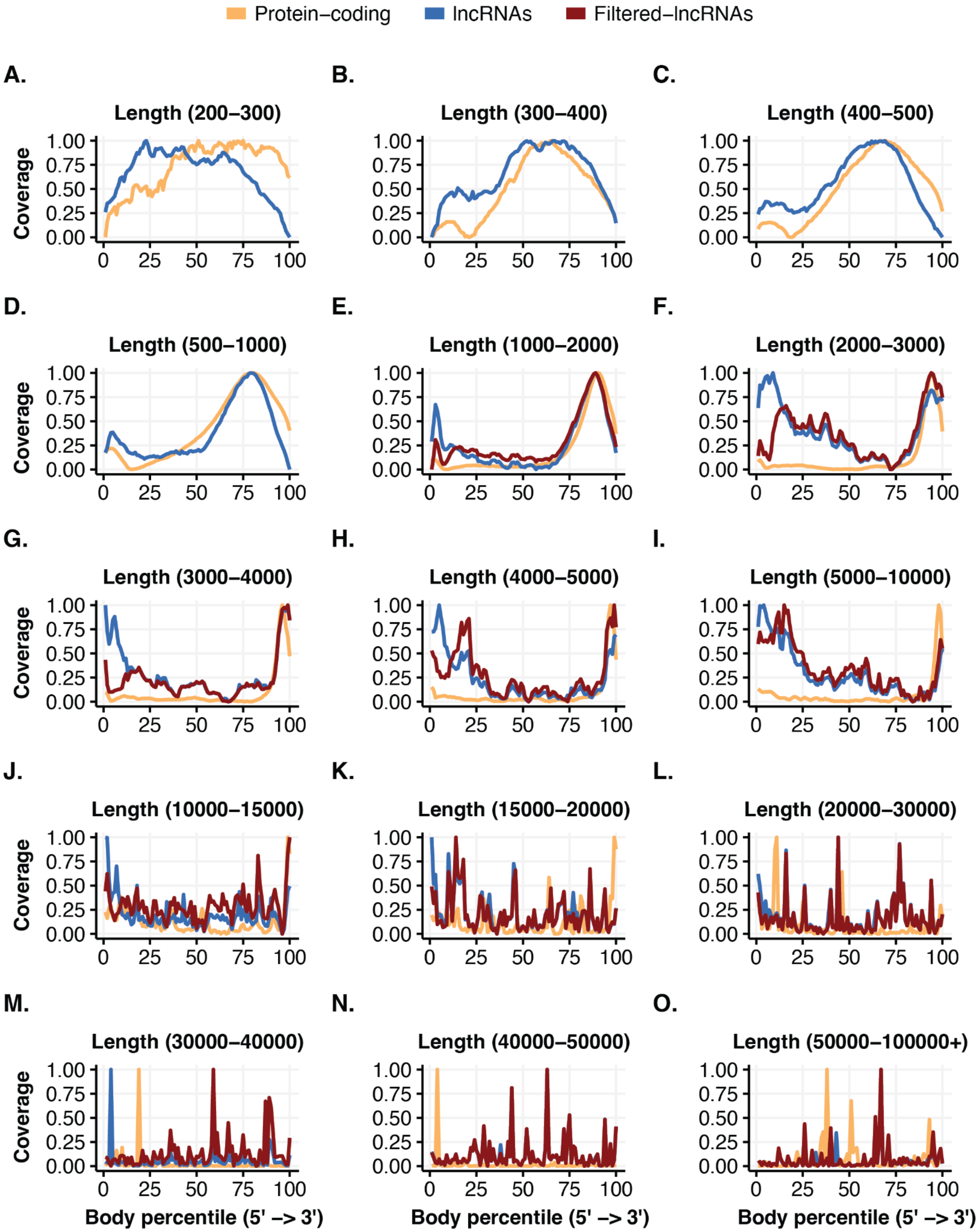
Supplementary Figure 4. Comparison of reads uniquely mapped to genome, transcriptome, intronic regions and intergenic regions with GENCODE, TLGA, and ULGA annotations. Y-axis shows percentage of reads mapped uniquely to **(A)** genome, **(B)** transcriptome, **(C)**, intronic regions and **(D)** intergenic regions with TLGA and ULGA compared to GENCODE in PBMCs, samples from liver set 1 (liver 1-5) (MacParland et al. 2018) and liver set 2 (liver 6-25) (Ramachandran et al. 2019).



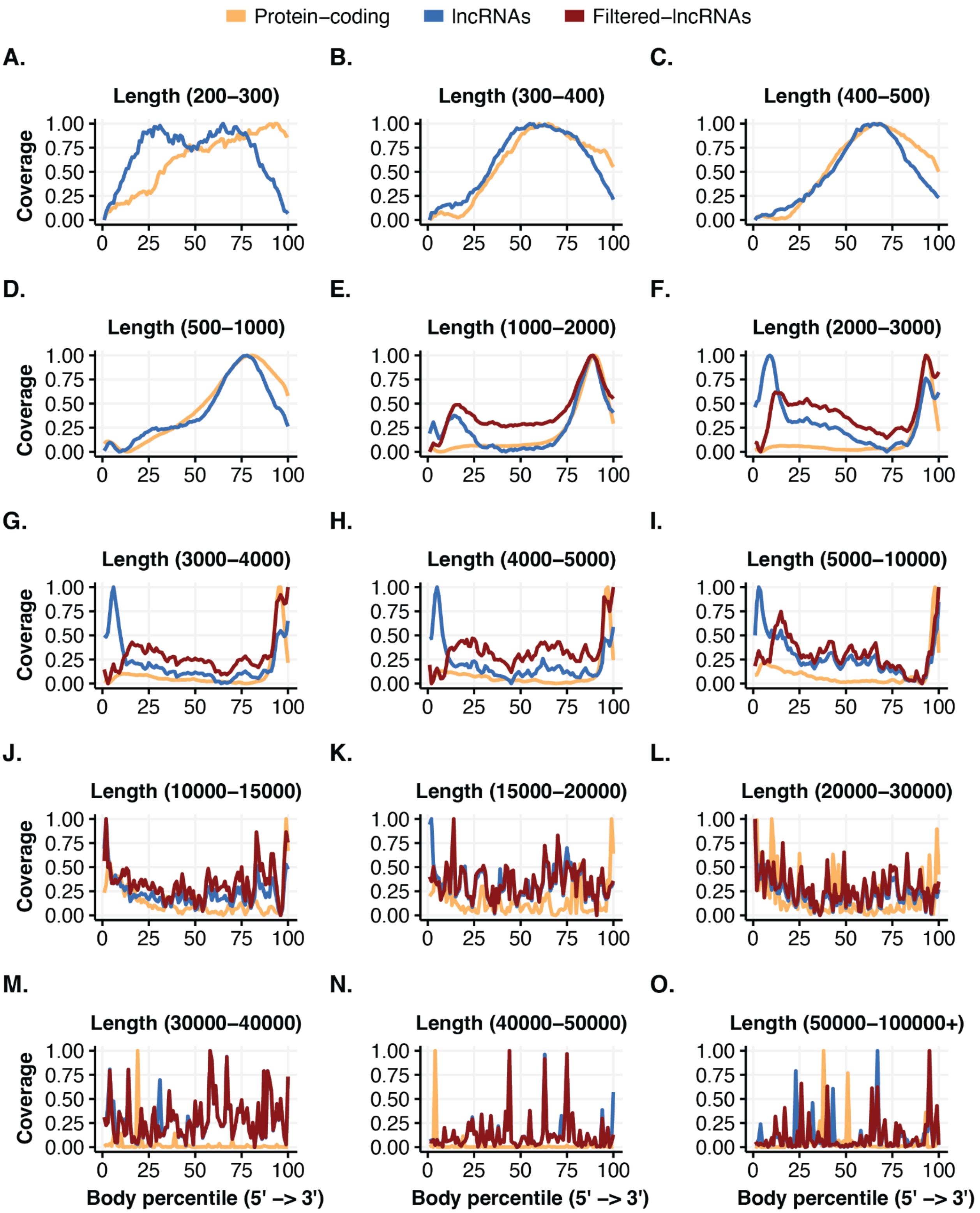
Supplementary Figure 5. Distribution of the mapped reads across the transcripts in liver. Distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue) in liver. The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates transcript coverage (0-1). **(A)** Expressed protein-coding genes (orange) and lncRNAs (blue) in liver set 1. **(B)** Expressed protein-coding genes (orange) and lncRNAs (blue) in liver set 2. In both liver set 1 and liver set 2 lncRNAs (blue) transcripts indicate an enrichment of reads at the 5' end of lncRNA transcripts.


Protein–coding
lncRNAs
Filtered–lncRNAs
A.
Length (200–300)
B.
Length (300–400)
C.
Length (400–500)
D.
Length (500–1000)
E.
Length (1000–2000)
F.
Length (2000–3000)
G.
Length (3000–4000)
H.
Length (4000–5000)
I.
Length (5000–10000)
J.
Length (10000–15000)
K.
Length (15000–20000)
L.
Length (20000–30000)
M.
Length (30000–40000)
N.
Length (40000–50000)
O.
Length (50000–100000+)
Coverage
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Body percentile (5' –> 3')

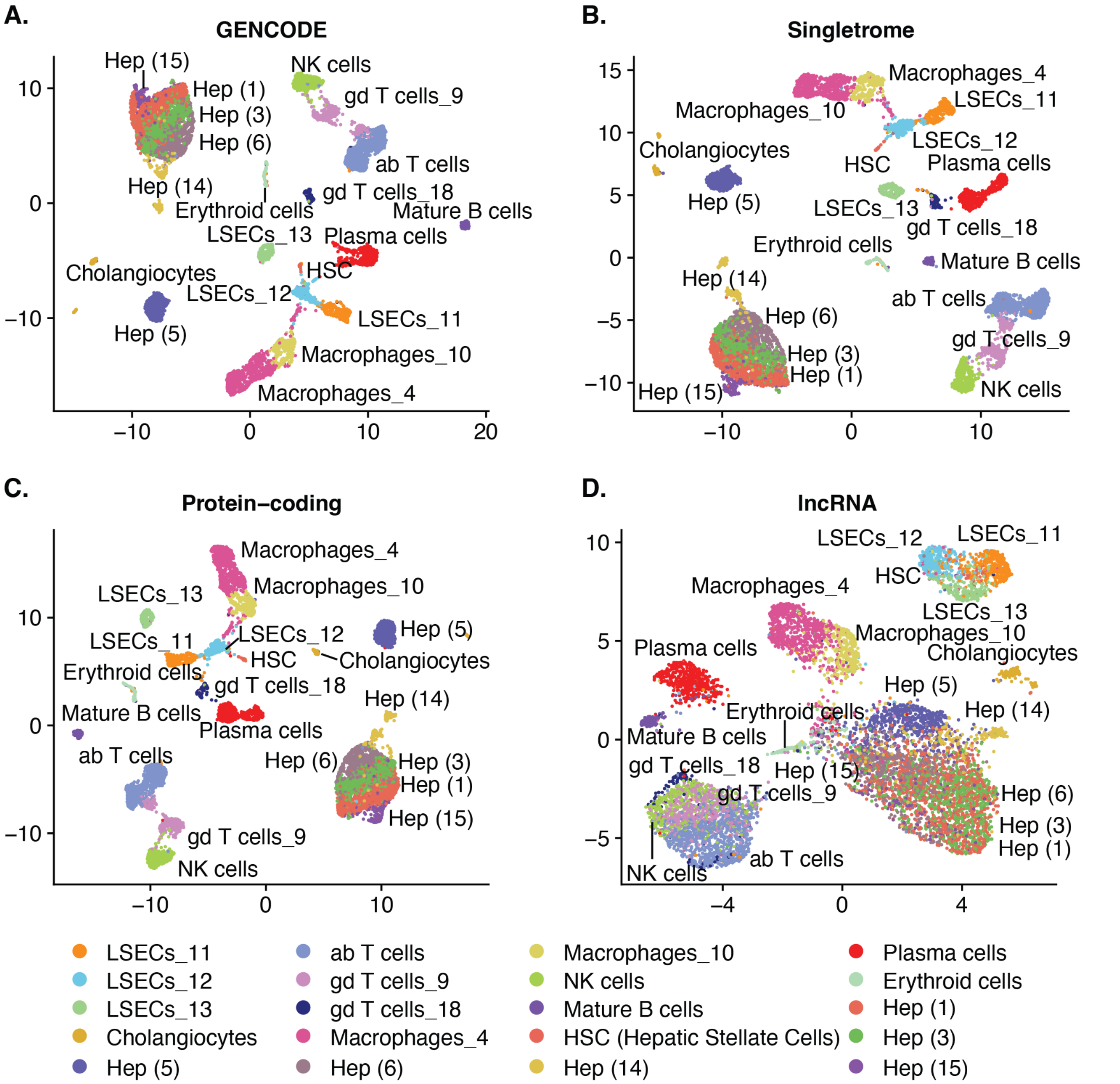
Supplementary Figure 6. Distribution of lncRNA mapping in PBMCs across transcripts of different lengths. (A-O), distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue) across different lengths of transcripts. Minimum and maximum length of the transcripts for each panel are shown at the top in parenthesis. The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates transcript coverage (0-1). lncRNA transcripts of 1000 or more nucleotides **(E-O)** were filtered, if reads mapped to a transcript indicate an enrichment of reads in the first 10 bins of lncRNA transcript, or if the majority of reads were mapped to a single location (1 bin) in the transcript and that location is in the first 90 bins. Filtered-lncRNAs (red line) shows the distribution of the mapped reads after removing lncRNAs that were flagged for low quality (material and methods).


Protein–coding
lncRNAs
Filtered–lncRNAs
A.
Length (200–300)
B.
Length (300–400)
C.
Length (400–500)
D.
Length (500–1000)
E.
Length (1000–2000)
F.
Length (2000–3000)
G.
Length (3000–4000)
H.
Length (4000–5000)
I.
Length (5000–10000)
J.
Length (10000–15000)
K.
Length (15000–20000)
L.
Length (20000–30000)
M.
Length (30000–40000)
N.
Length (40000–50000)
O.
Length (50000–100000+)
Coverage
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Body percentile (5' –> 3')

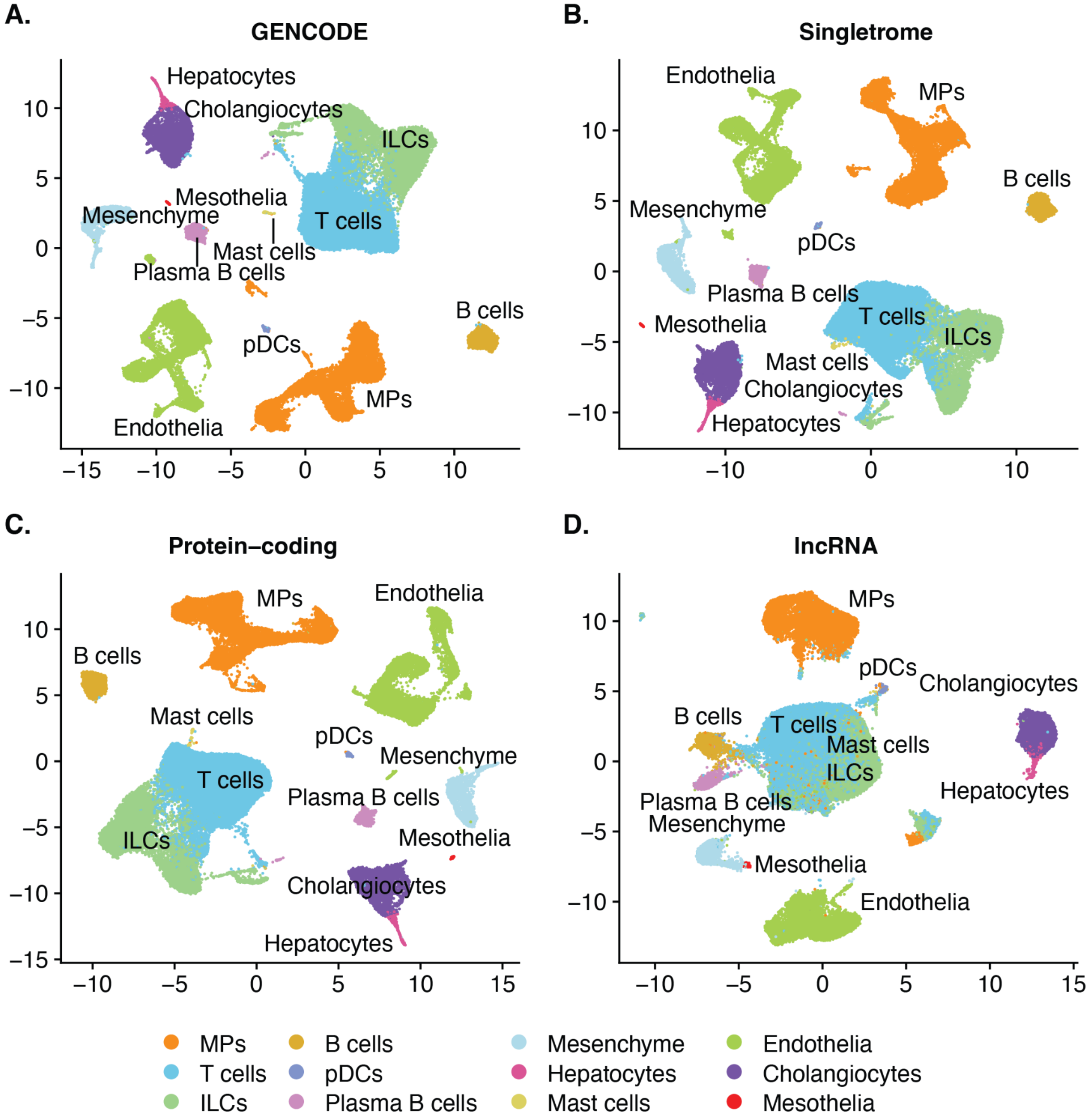
Supplementary Figure 7. Distribution of lncRNA mapping in liver set 1 across transcripts of different lengths. (A-O), distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue) across different lengths of transcripts. Minimum and maximum length of the transcripts for each panel are shown at the top in parenthesis. The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates transcript coverage (0-1). lncRNA transcripts of 1000 or more nucleotides **(E-O)** were filtered, if reads mapped to a transcript indicate an enrichment of reads in the first 10 bins of lncRNA transcript, or if the majority of reads were mapped to a single location (1 bin) in the transcript and that location is in the first 90 bins. Filtered-lncRNAs (red line) shows the distribution of the mapped reads after removing lncRNAs that were flagged for low quality (material and methods).


Protein−coding
lncRNAs
Filtered−lncRNAs
A.
Length (200−300)
B.
Length (300−400)
C.
Length (400−500)
D.
Length (500−1000)
E.
Length (1000−2000)
F.
Length (2000−3000)
G.
Length (3000−4000)
H.
Length (4000−5000)
I.
Length (5000−10000)
J.
Length (10000−15000)
K.
Length (15000−20000)
L.
Length (20000−30000)
M.
Length (30000−40000)
N.
Length (40000−50000)
O.
Length (50000−100000+)
Coverage
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Body percentile (5' −> 3')

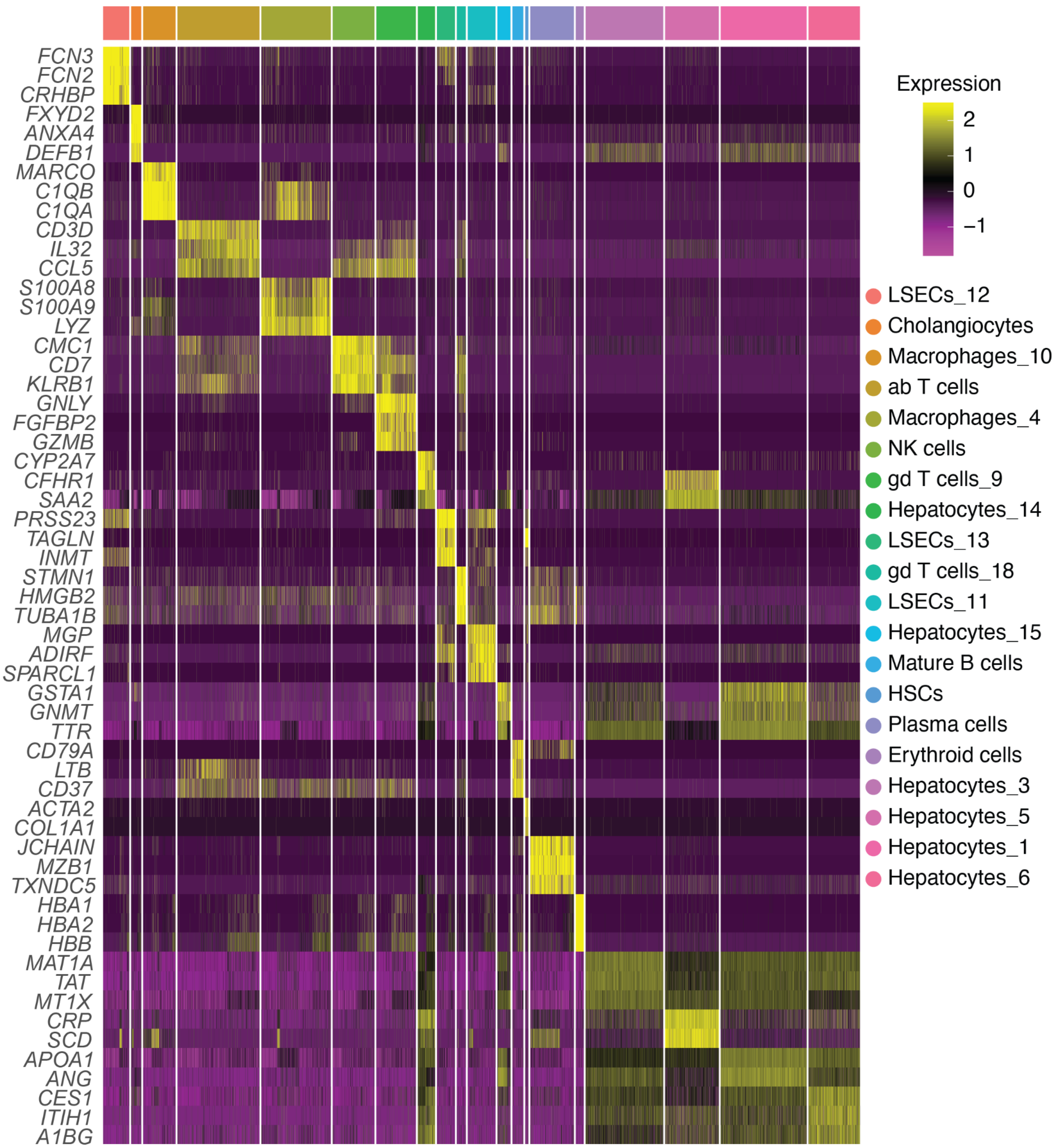
Supplementary Figure 8. Distribution of lncRNA mapping in liver set 2 across transcripts of different lengths. (A-O), distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue) across different lengths of transcripts. Minimum and maximum length of the transcripts for each panel are shown at the top in parenthesis. The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates transcript coverage (0-1). lncRNA transcripts of 1000 or more nucleotides **(E-O)** were filtered, if reads mapped to a transcript indicate an enrichment of reads in the first 10 bins of lncRNA transcript, or if the majority of reads were mapped to a single location (1 bin) in the transcript and that location is in the first 90 bins. Filtered-lncRNAs (red line) shows the distribution of the mapped reads after removing lncRNAs that were flagged for low quality (material and methods).


A.
GENCODE
Hep (15)
Hep (1)
Hep (3)
Hep (6)
Hep (14)
Erythroid cells
NK cells
gd T cells_9
ab T cells
gd T cells_18
Mature B cells
LSECs_13
Plasma cells
Cholangiocytes
HSC
LSECs_12
LSECs_11
Hep (5)
Macrophages_10
Macrophages_4
10
0
−10
−10
0
10
20
B.
Singletrome
Macrophages_4
LSECs_11
Macrophages_10
LSECs_12
Cholangiocytes
HSC
Plasma cells
Hep (5)
LSECs_13
gd T cells_18
Erythroid cells
Mature B cells
Hep (14)
ab T cells
Hep (6)
gd T cells_9
Hep (3)
Hep (1)
Hep (15)
NK cells
15
10
5
0
−5
−10
−10
0
10
C.
Protein−coding
Macrophages_4
Macrophages_10
LSECs_13
LSECs_12
Hep (5)
LSECs_11
HSC
Cholangiocytes
Erythroid cells
gd T cells_18
Hep (14)
Mature B cells
Plasma cells
ab T cells
Hep (6)
Hep (3)
Hep (1)
Hep (15)
gd T cells_9
NK cells
10
0
−10
−10
0
10
D.
lncRNA
LSECs_12
LSECs_11
HSC
Macrophages_4
LSECs_13
Macrophages_10
Plasma cells
Cholangiocytes
Hep (5)
Erythroid cells
Hep (14)
Mature B cells
gd T cells_18
Hep (15)
gd T cells_9
Hep (6)
Hep (3)
Hep (1)
ab T cells
NK cells
10
5
0
−5
−4
0
4
● LSECs_11
● LSECs_12
● LSECs_13
● Cholangiocytes
● Hep (5)
● ab T cells
● gd T cells_9
● gd T cells_18
● Macrophages_4
● Hep (6)
● Macrophages_10
● NK cells
● Mature B cells
● HSC (Hepatic Stellate Cells)
● Hep (14)
● Plasma cells
● Erythroid cells
● Hep (1)
● Hep (3)
● Hep (15)

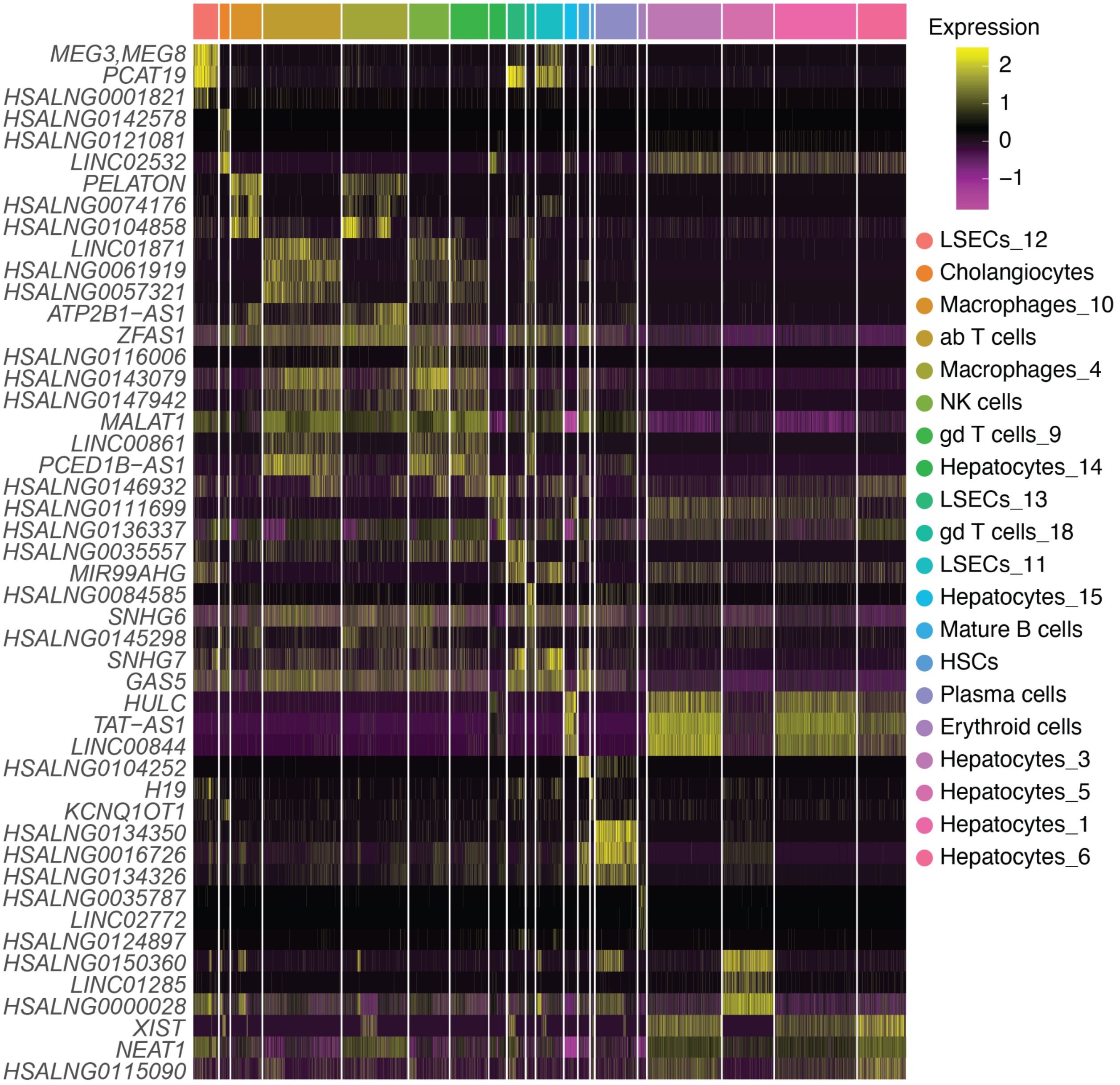
Supplementary Figure 9. lncRNAs alone predict most clusters and cell types in single cell data of liver set 1. scRNA-seq data of liver set 1 were mapped using annotation from **(A)** GENCODE, **(B)** Singletrome, **(C)** only protein-coding genes in Singletrome, and **(D)** only lncRNAs in Singletrome. The labels for each cell were retained from the original publications. For this analysis, Singletrome only contains lncRNAs that meet all filters developed with analysis of PBMCs and applied to data from liver set 1. Hepatocytes are abbreviated as Hep and hepatic stellate cells are abbreviated as HSC.


A.
GENCODE
Hepatocytes
Cholangiocytes
ILCs
Mesothelia
Mesenchyme
T cells
Mast cells
Plasma B cells
B cells
pDCs
MPs
Endothelia
10
5
0
−5
−10
−15
−10
−5
0
5
10
B.
Singletrome
Endothelia
MPs
B cells
Mesenchyme
pDCs
Plasma B cells
T cells
Mesothelia
ILCs
Mast cells
Cholangiocytes
Hepatocytes
10
5
0
−5
−10
−10
0
10
C.
Protein−coding
MPs
Endothelia
B cells
Mast cells
pDCs
Mesenchyme
T cells
Plasma B cells
ILCs
Mesothelia
Cholangiocytes
Hepatocytes
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
D.
lncRNA
MPs
pDCs
Cholangiocytes
B cells
T cells
Mast cells
ILCs
Plasma B cells
Hepatocytes
Mesenchyme
Mesothelia
Endothelia
10
5
0
−5
−10
−10
−5
0
5
10
15
● MPs
● T cells
● ILCs
● B cells
● pDCs
● Plasma B cells
● Mesenchyme
● Hepatocytes
● Mast cells
● Endothelia
● Cholangiocytes
● Mesothelia

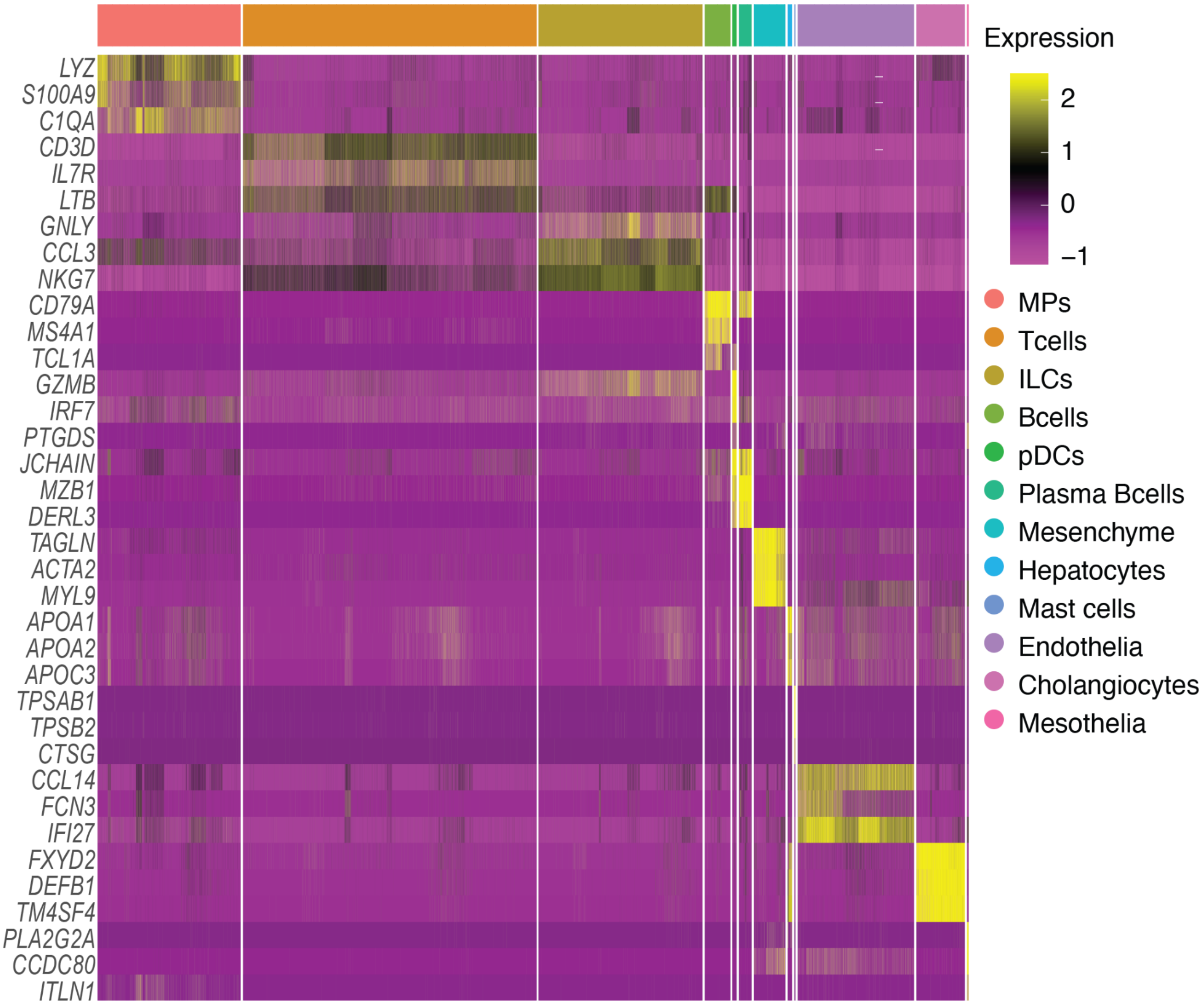
Supplementary Figure 10. lncRNAs alone predict most clusters and cell types in single cell data of liver set 2. scRNA-seq data of liver set 2 were mapped using annotation from **(A)** GENCODE, **(B)** Singletrome, **(C)** only protein-coding genes in Singletrome, and **(D)** only lncRNAs in Singletrome. The labels for each cell were retained from the original publications. For this analysis, Singletrome only contains lncRNAs that meet all filters developed with analysis of PBMCs and applied to data from liver set 2.



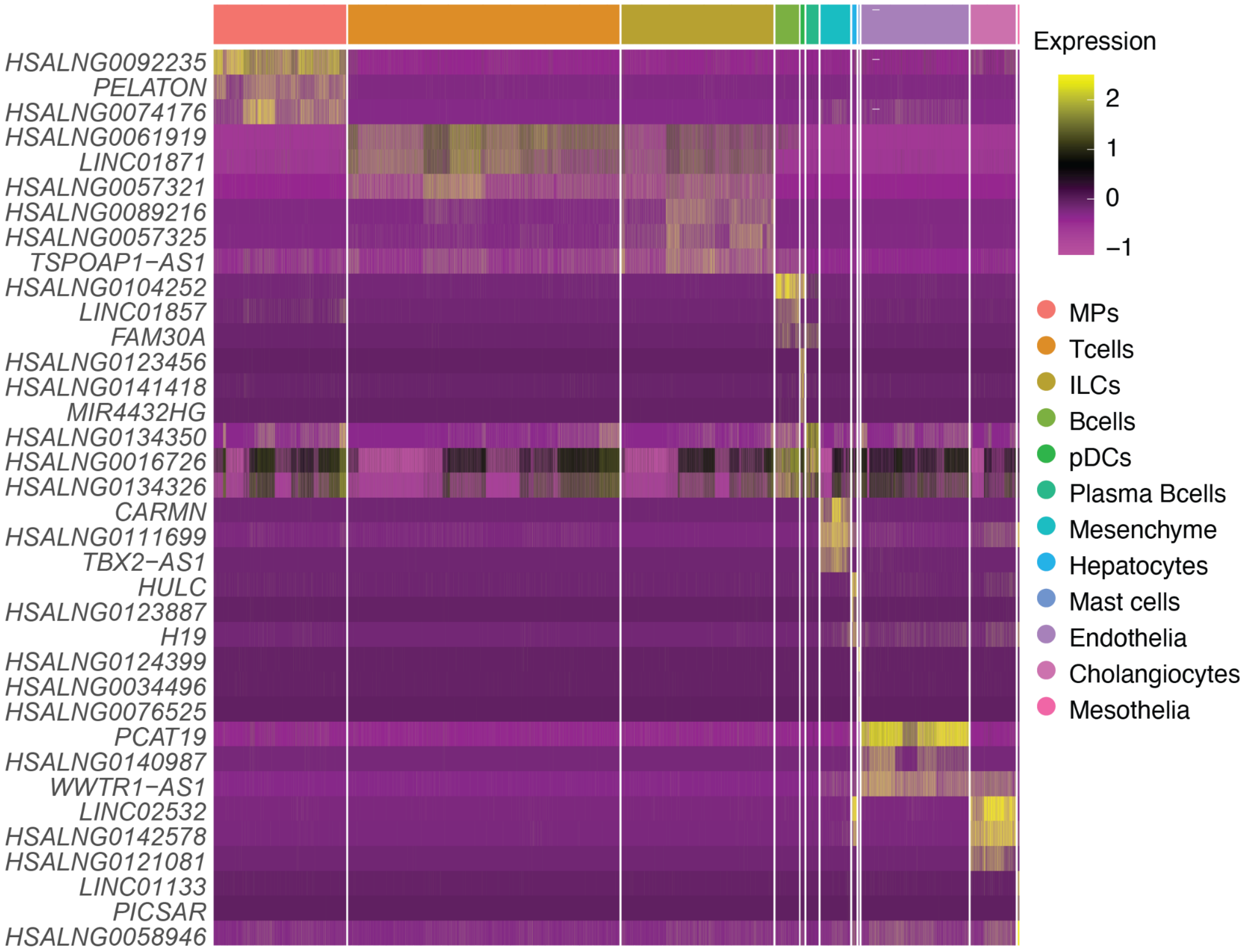
Supplementary Figure 11. Protein-coding based cell type markers for liver set 1. The heatmap displays the top differentially expressed protein-coding genes (y-axis) for each cell type in liver set 1. Cell types are indicated by color at the bar above the heatmap, and the key is displayed to the right. Expression level is indicated by Z-score.



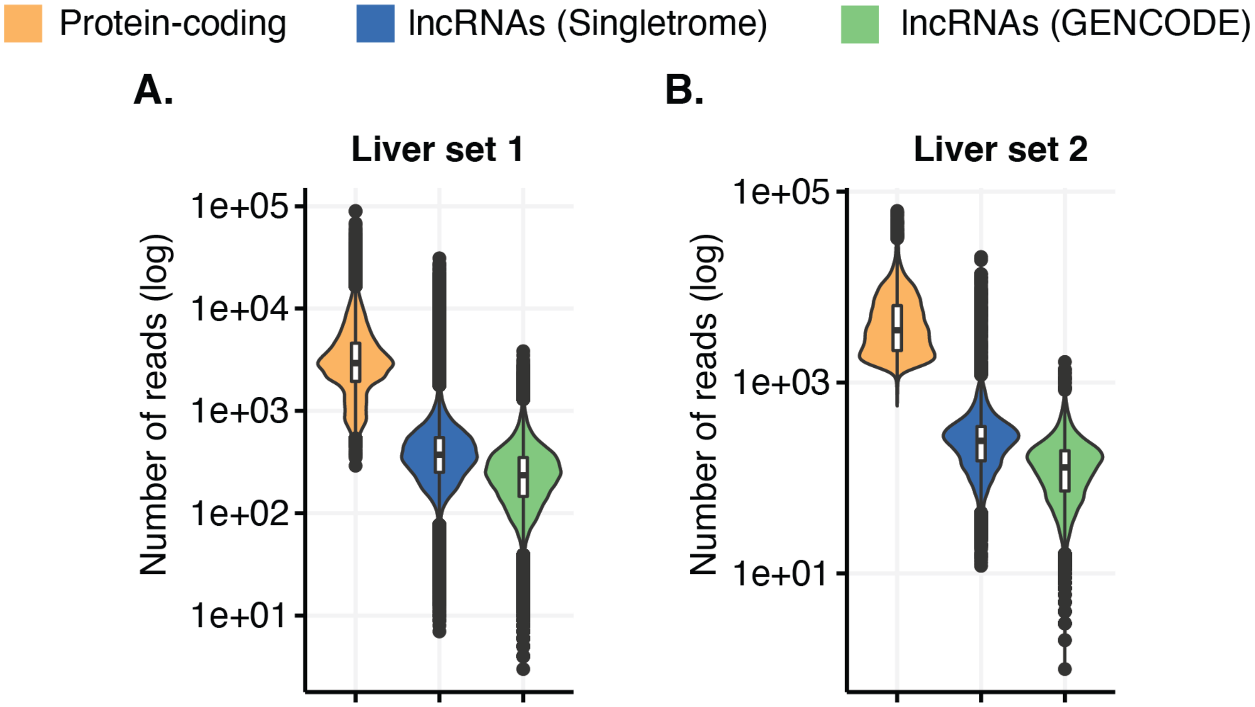
Supplementary Figure 12. lncRNA based cell type markers for liver set 1. The heatmap displays the top differentially expressed lncRNA genes (y-axis) for each cell type in liver set 1. Cell types are indicated by color at the bar above the heatmap, and the key is displayed to the right. Expression level is indicated by Z-score.



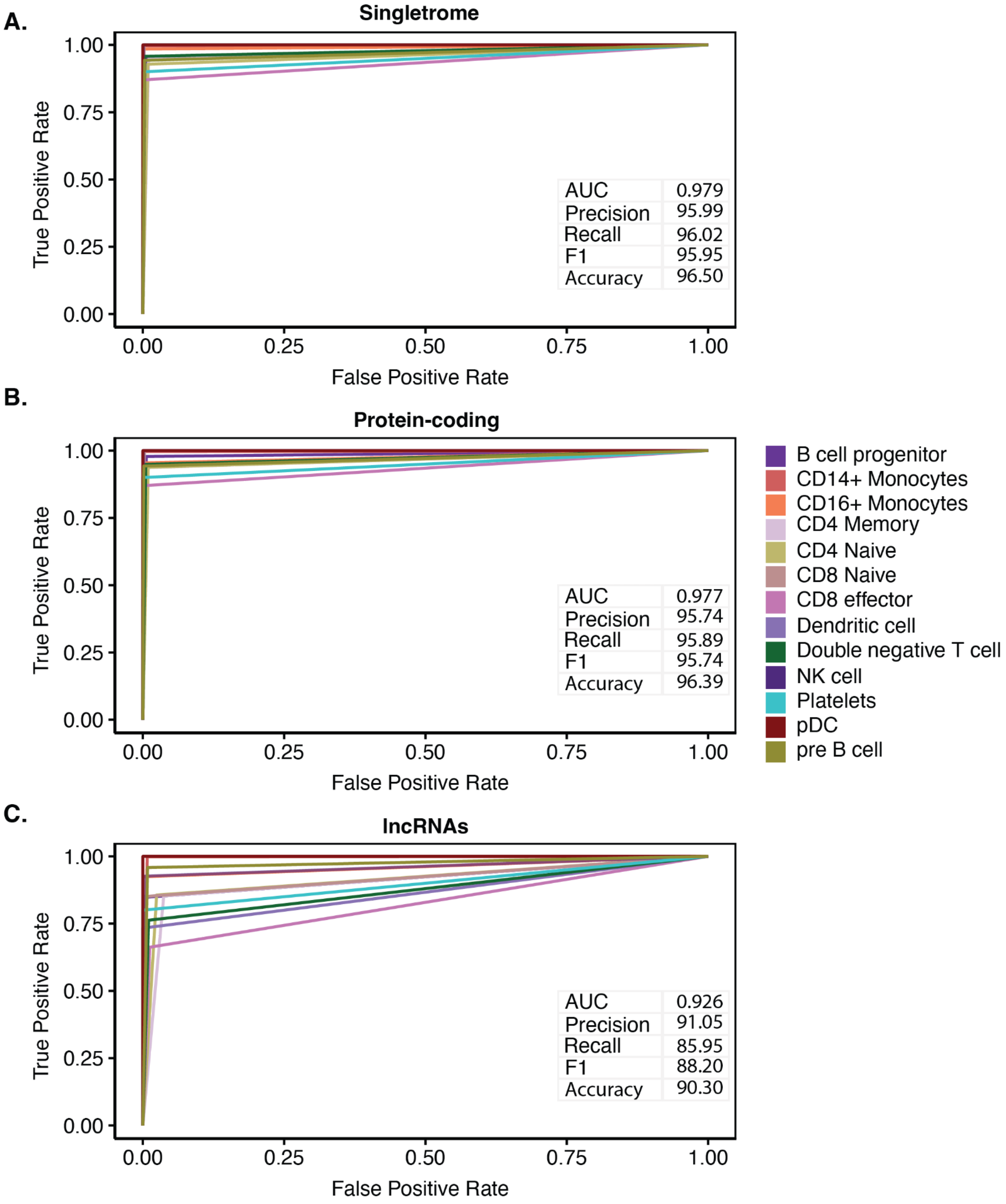
Supplementary Figure 13. Protein-coding based cell type markers for liver set 2. The heatmap displays the top differentially expressed protein-coding genes (y-axis) for each cell type in liver set 2. Cell types are indicated by color at the bar above the heatmap, and the key is displayed to the right. Expression level is indicated by Z-score.



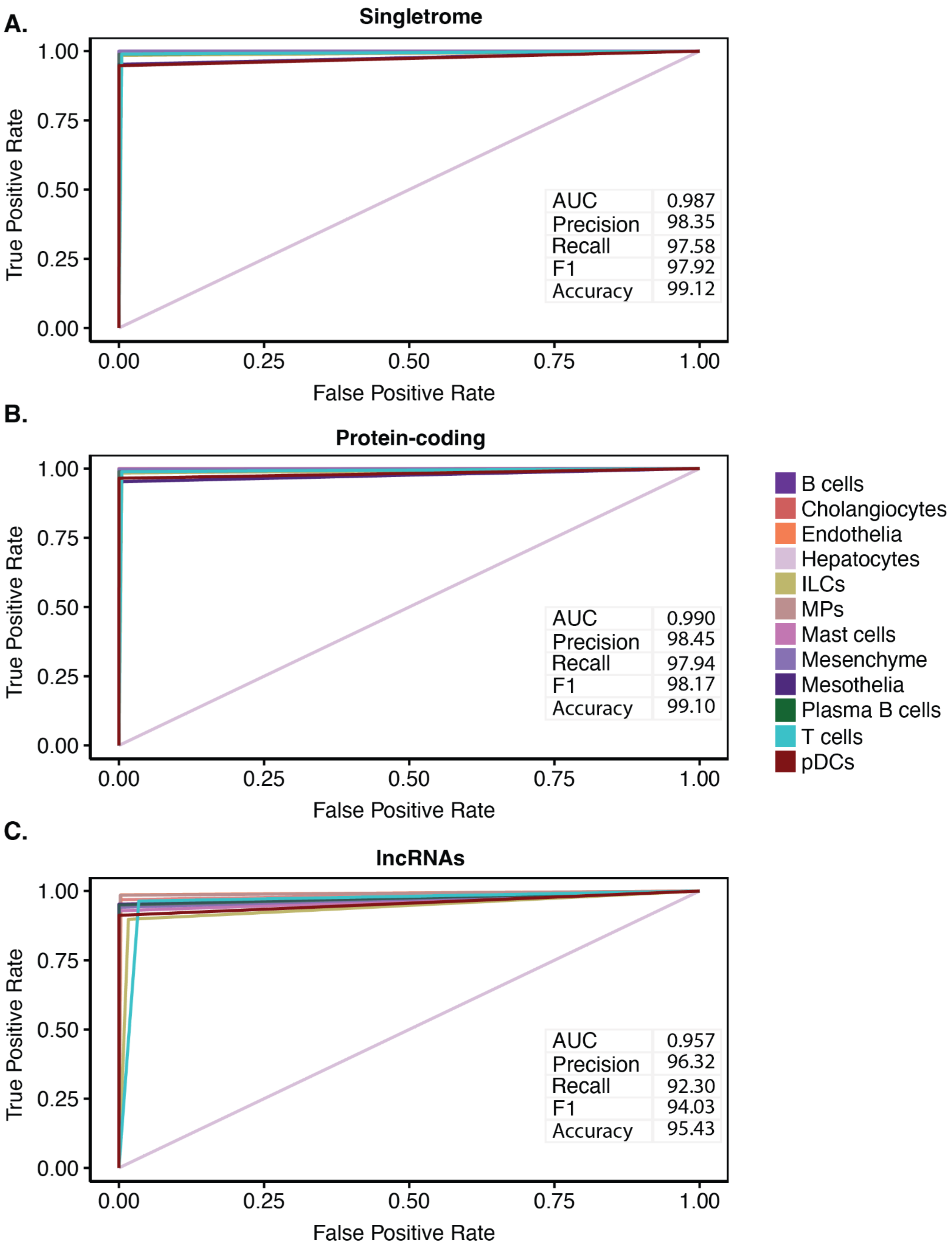
Supplementary Figure 14. lncRNA based cell type markers for liver set 2. The heatmap displays the top differentially expressed lncRNA genes (y-axis) for each cell type in liver set 2. Cell types are indicated by color at the bar above the heatmap, and the key is displayed to the right. Expression level is indicated by Z-score.



Supplementary Figure 15. Expression of quality filtered lncRNAs compared to protein-coding genes in liver. The total number of mapped reads per cell (y-axis, log scale) is quantified for protein-coding genes (orange), lncRNA genes from Singletrome (blue), and lncRNA genes from GENCODE (green) in liver set 1 **(A)** and liver set 2 **(B).**



Supplementary Figure 16. Cell type prediction for PBMCs. Receiver-operating characteristic (ROC) curve showing true and false positive rates for cell type prediction based on the expression of all genes in Singletrome **(A),** protein-coding genes alone **(B),** and lncRNA genes alone **(C)**. Cell types are indicated by color of the line, and the key is displayed to the right. The table inside the panel of each (A-C) shows the AUC, precision (%), recall (%), F1 (%), and accuracy (%) for cell type prediction of PBMCs.

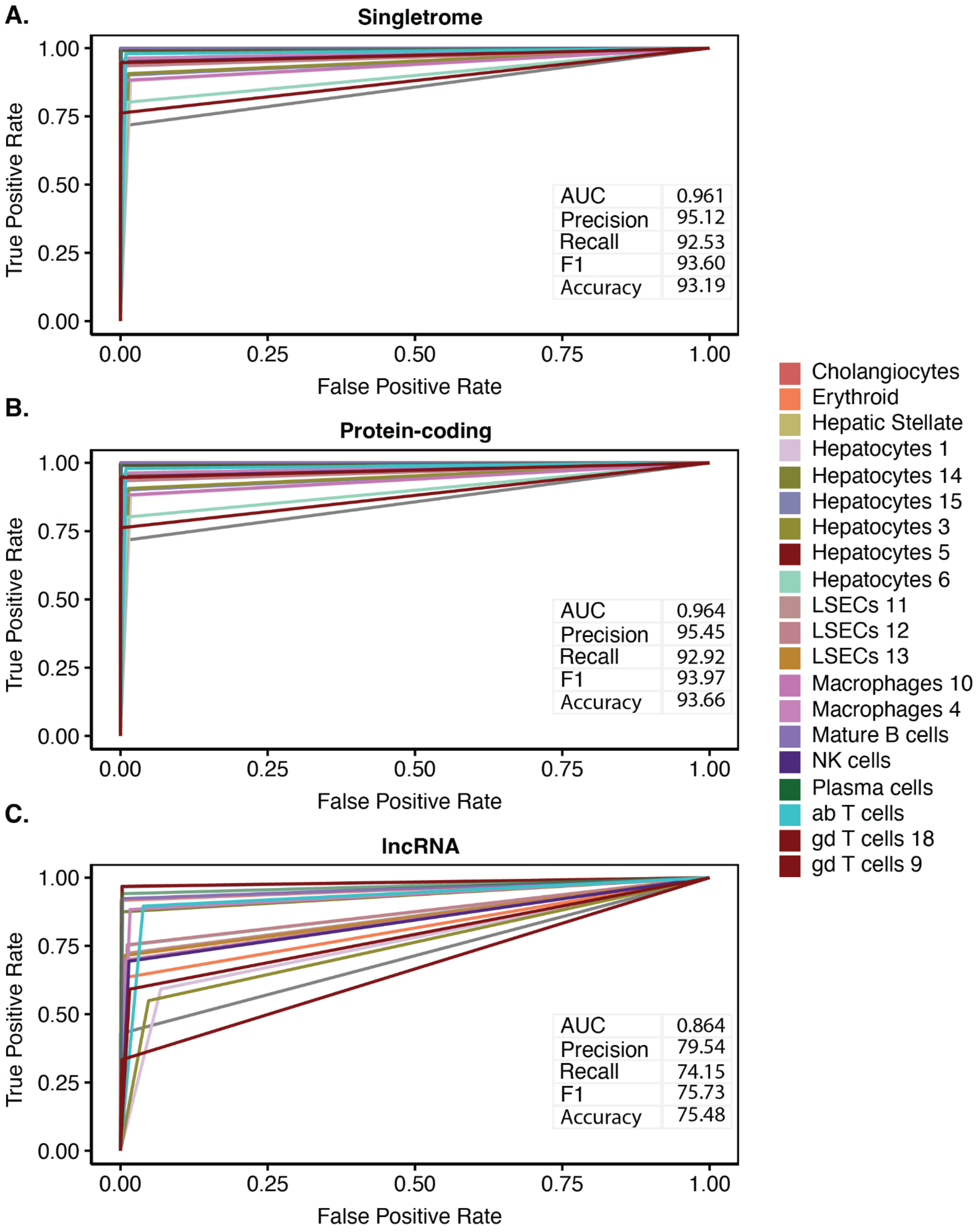


AUC	0.987
Precision	98.35
Recall	97.58
F1	97.92
Accuracy	99.12

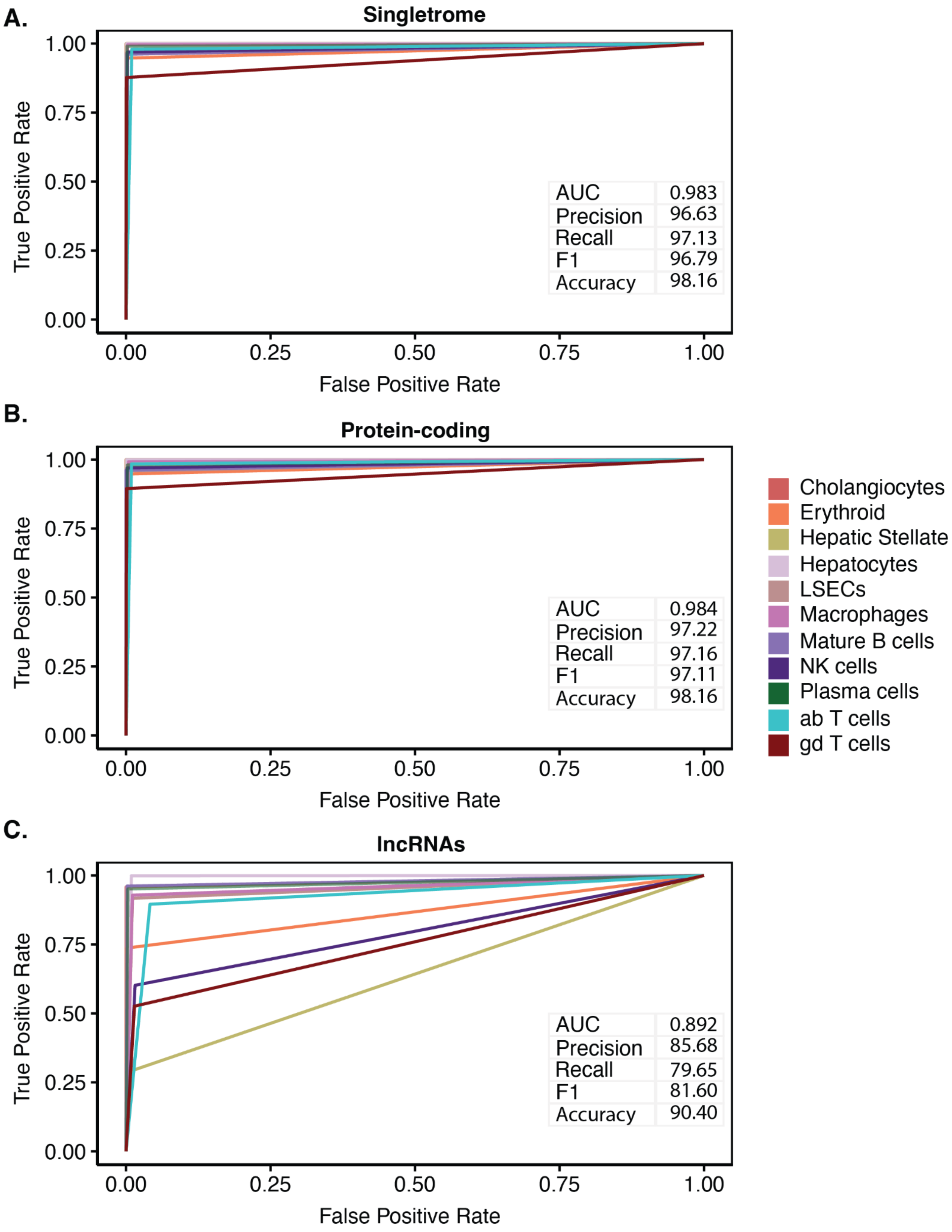
AUC	0.990
Precision	98.45
Recall	97.94
F1	98.17
Accuracy	99.10

AUC	0.957
Precision	96.32
Recall	92.30
F1	94.03
Accuracy	95.43

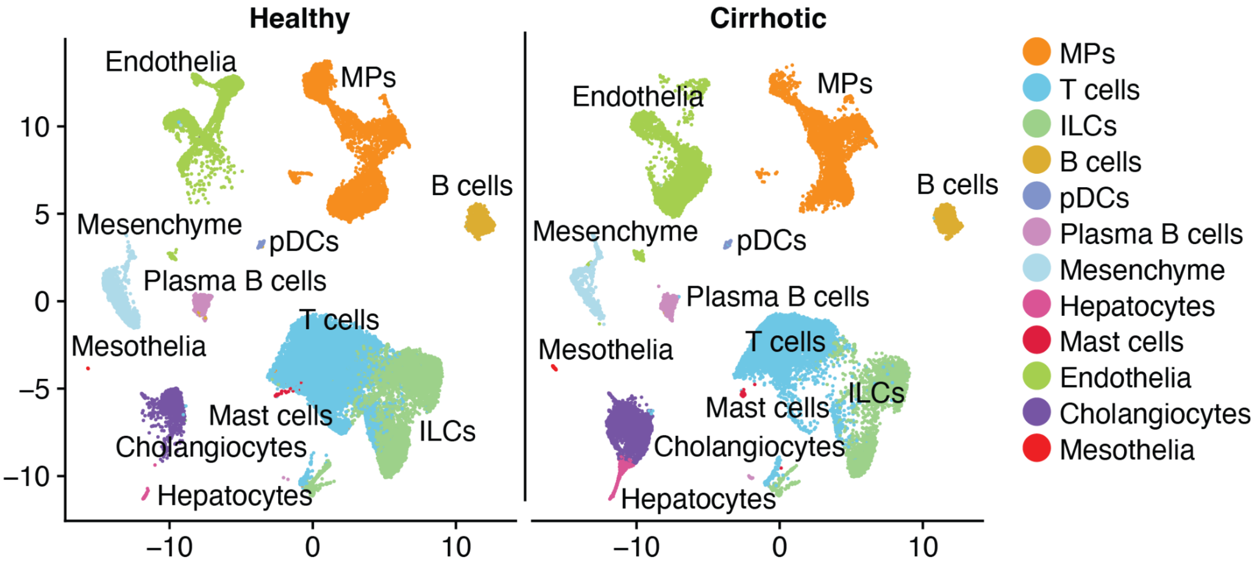
Supplementary Figure 17. Cell type prediction for liver set 2. Receiver-operating characteristic (ROC) curve showing true and false positive rates for cell type prediction based on the expression of all genes in Singletrome **(A),** protein-coding genes alone **(B),** and lncRNA genes alone **(C)**. Cell types are indicated by color of the line, and the key is displayed to the right. The table inside the panel of each (A-C) shows the AUC, precision (%), recall (%), F1 (%), and accuracy (%) for cell type prediction of liver set 2.


A.
Singletrome
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC 0.961
Precision 95.12
Recall 92.53
F1 93.60
Accuracy 93.19
0.00
0.25
0.50
0.75
1.00
False Positive Rate
B.
Protein-coding
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC 0.964
Precision 95.45
Recall 92.92
F1 93.97
Accuracy 93.66
0.00
0.25
0.50
0.75
1.00
False Positive Rate
C.
lncRNA
True Positive Rate
1.00
0.75
0.50
0.25
0.00
AUC 0.864
Precision 79.54
Recall 74.15
F1 75.73
Accuracy 75.48
0.00
0.25
0.50
0.75
1.00
False Positive Rate
Cholangiocytes
Erythroid
Hepatic Stellate
Hepatocytes 1
Hepatocytes 14
Hepatocytes 15
Hepatocytes 3
Hepatocytes 5
Hepatocytes 6
LSECs 11
LSECs 12
LSECs 13
Macrophages 10
Macrophages 4
Mature B cells
NK cells
Plasma cells
ab T cells
gd T cells 18
gd T cells 9

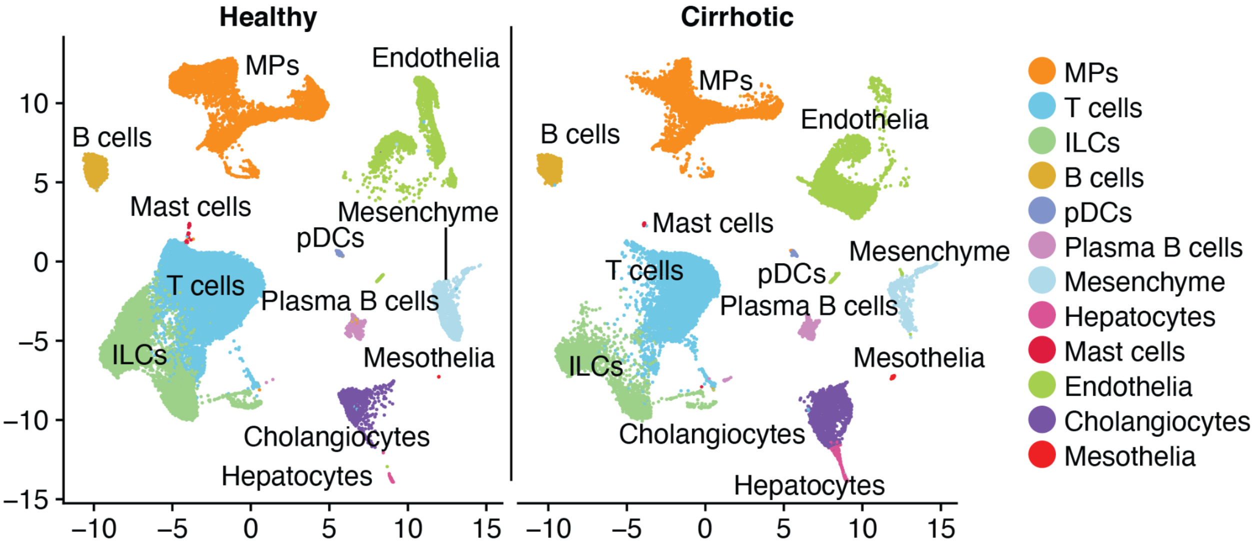
Supplementary Figure 18. Cell type prediction for liver set 1. Receiver-operating characteristic (ROC) curve showing true and false positive rates for cell type prediction based on the expression of all genes in Singletrome **(A),** protein-coding genes alone **(B),** and lncRNA genes alone **(C)**. Cell types are indicated by color of the line, and the key is displayed to the right. The table inside the panel of each (A-C) shows the AUC, precision (%), recall (%), F1 (%), and accuracy (%) for cell type prediction of liver set 1.


A.
Singletrome
True Positive Rate
1.00
0.75
0.50
0.25
0.00
False Positive Rate
0.00
0.25
0.50
0.75
1.00
AUC 0.983
Precision 96.63
Recall 97.13
F1 96.79
Accuracy 98.16
B.
Protein-coding
True Positive Rate
False Positive Rate
AUC 0.984
Precision 97.22
Recall 97.16
F1 97.11
Accuracy 98.16
C.
lncRNAs
True Positive Rate
False Positive Rate
AUC 0.892
Precision 85.68
Recall 79.65
F1 81.60
Accuracy 90.40
Cholangiocytes
Erythroid
Hepatic Stellate
Hepatocytes
LSECs
Macrophages
Mature B cells
NK cells
Plasma cells
ab T cells
gd T cells

Supplementary Figure 19. Cell type prediction for liver set 1 (sub-clusters merged to the same cell type). Receiver-operating characteristic (ROC) curve showing true and false positive rates for cell type prediction based on the expression of all genes in Singletrome **(A),** protein-coding genes alone **(B),** and lncRNA genes alone **(C)**. Cell types are indicated by color of the line, and the key is displayed to the right. The table inside the panel of each (A-C) shows the AUC, precision (%), recall (%), F1 (%), and accuracy (%) for cell type prediction of liver set 1.



Supplementary Figure 20. Singletrome cell type map in healthy and cirrhotic liver. scRNA-seq data of liver set 2 (GSE136103(Ramachandran et al. 2019)) were mapped using annotations from Singletrome. The labels for each cell were retained from the original publication. Cells were clustered based on all the genes in Singletrome and annotated by condition healthy (left) and cirrhotic liver (right). For this analysis, Singletrome contains all the protein-coding genes and only lncRNAs that meet all the described filters in the section ‘Quality control of lncRNA mapping’.



Supplementary Figure 21. Protein-coding cell type map in healthy and cirrhotic liver. scRNA-seq data of liver set 2 (GSE136103(Ramachandran et al. 2019)) were mapped using annotations from Singletrome. The labels for each cell were retained from the original publication. Cells were clustered based on protein-coding genes from Singletrome and annotated by condition healthy (left) and cirrhotic liver (right).

Supplementary material

	PBMCs	Liver set 1	Liver set 2
Total expressed genes in ULGA	39599	36262	57210
Total expressed lncRNAs in ULGA	25470	20813	40375
Total expressed lncRNAs in TLGA	24034	19692	38576
Overlapping antisense lncRNAs expressed in ULGA	6379	5347	8516
Overlapping antisense lncRNAs expressed in TLGA	4924	4201	6681
Overlapping antisense lncRNAs expressed only in ULGA	1458	1153	1841
Overlapping antisense lncRNAs expressed in both TLGA and ULGA	4921	4194	6675
Median expression of common (antisense overlapping) lncRNAs in ULGA	174	45	70
Median expression of common (antisense overlapping) lncRNAs in TLGA	142	38	57
ULGA Genes after filtering for TLGA	24012	19660	38534
Singletrome:Total Genes after filtering	38141	35109	55369

Supplementary Table 1. Total detected genes in TLGA and ULGA. The table shows a comparison of the antisense overlapping lncRNA genes that are expressed commonly in ULGA and TLGA or expressed only in ULGA. The median of the read count for the lncRNA genes expressed commonly in TLGA is lower than the ULGA.

	GENCODE	TLGA	ULGA
PBMCs	5064	24034	25470
Liver set 1	4800	19692	20813
Liver set 2	8211	38576	40375

Supplementary Table 2. Number of expressed lncRNA genes in GENCODE, TLGA and ULGA. ULGA detected significantly more lncRNAs compared to GENCODE (the most widely used genome annotation for scRNA-seq analysis). lncRNAs are considered as expressed and utilized for the down-stream analysis if they are expressed in at least 10 cells in a dataset (similar to protein-coding genes).

Min transcript length	Max transcript length	Number of protein-coding transcripts	Number of lncRNA transcripts
200	300	1721	19279
300	400	4018	19877
400	500	9120	21362
500	1000	61852	86054
1000	2000	27848	81517
2000	3000	19481	35542
3000	4000	10793	18395
4000	5000	6314	10355
5000	10000	8629	15126
10000	15000	923	2471
15000	20000	158	567
20000	30000	48	286
30000	40000	3	59
40000	50000	2	27
50000	10000+	7	21

Supplementary Table 3. Number of protein-coding and lncRNA transcripts in different length ranges. These sets of transcripts were used to calculate read coverage across the transcript body of protein-coding and lncRNA genes.

Minimum Transcript length	Maximum Transcript length	Protein-coding correlation	lncRNA correlation
200	300	0.28	0.22
300	400	0.28	0.22
400	500	0.27	0.2
500	1000	0.3	0.21
1000	2000	0.3	0.17
2000	3000	0.27	0.11
3000	4000	0.19	0.07
4000	5000	0.15	0.03
5000	10000	0.09	-0.01
10000	15000	0.04	-0.08
15000	20000	0.37	-0.12
20000	30000	0.06	-0.08
30000	40000	0.07	-0.01
40000	50000	-0.23	-0.02
50000	100000+	NA	-0.04

Supplementary Table 4. Correlation between transcript length and gene length for protein-coding genes and lncRNA genes. Correlation was calculated for all transcripts of a gene if it contained at least one transcript in the length range of minimum transcript length and maximum transcript length.

	PBMCs	Liver set 1	Liver set 2
5' high expressed lncRNA genes	2445	3065	4486
5' high expressed lncRNA genes discarded	433	488	928
5' high expressed lncRNA transcripts discarded	5685	7372	9296
Peak bin expression lncRNA genes	606	644	1084
Peak bin lncRNA genes discarded	67	45	98
Peak bin lncRNA transcripts discarded	1455	1312	2271
Total expressed lncRNAs in ULGA	25470	20813	40375
Total retained lncRNAs after QC filtering	23510	19126	37507

Supplementary Table 5. Quality control of lncRNA genes**.** lncRNA transcripts and genes were discarded. if lncRNA mapped reads exhibit 5' bias in 3' sequenced scRNA-seq libraries or if the majority of reads were mapped to a single location (peak bin expression) in the transcript, as both situations could represent library artifacts or mapping anomalies (Ma and Kingsford 2019). LncRNA genes for which all transcripts met either criterion in a dataset were excluded from further analysis in that dataset.

Biotype	Number of Genes
lncRNA	16562
protein_coding	19394
IG_V_pseudogene	188
IG_V_gene	144
IG_C_gene	14
IG_J_gene	18
TR_C_gene	6
TR_J_gene	79
TR_V_gene	106
TR_V_pseudogene	33
TR_D_gene	4
IG_C_pseudogene	9
TR_J_pseudogene	4
IG_J_pseudogene	3
IG_D_gene	37

Supplementary Table 6. GENCODE v32 (Frankish et al. 2019) human genome (dated 27.10.2021) number of genes per biotype. This genome annotation file was downloaded from https://cf.10xgenomics.com/supp/cell-exp/refdata-gex-GRCh38-2020-A.tar.gz. Of note, 10x genomics have filtered the GENCODE v32 to contain genes and transcripts only with the above-mentioned biotypes.

MKREF COUNT	CRv3.1	CRv4.0	CRv5.0.1	CRv6.0
CRv3.1	True	True	True	True
CRv4.0	False	True	True	True
CRv5.0.1	False	False	True	True
CRv6.0	False	False	True	True

Supplementary Table 7. The reference compatibility table for custom references from 10x genomics. Cell Ranger v3.1 mkref is stable across all count pipelines and it is compatible with the v1 chemistry.