

eScholarship@UMassChan

Singletrome enhances detection of long noncoding RNAs in single cell transcriptomes

Item Type	Journal Article
Authors	Rahman, Raza Ur;Ahmad, Iftikhar;Li, Zixiu;Sparks, Robert P;Ben Saad, Amel;Mullen, Alan C
Citation	Rahman RU, Ahmad I, Li Z, Sparks RP, Ben Saad A, Mullen AC. Singletrome enhances detection of long noncoding RNAs in single cell transcriptomes. Sci Rep. 2025 Aug 12;15(1):29542. doi: 10.1038/s41598-025-13528-9. PMID: 40796606; PMCID: PMC12344142.
DOI	10.1038/s41598-025-13528-9
Rights	Open Access: This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit https://creativecommons.org/licenses/by-nc-nd/4.0/ . © The Author(s) 2025
Download date	2026-03-16 20:50:33
Item License	http://creativecommons.org/licenses/by-nc-nd/4.0/
Link to Item	https://hdl.handle.net/20.500.14038/54755



OPEN Singletrome enhances detection of long noncoding RNAs in single cell transcriptomes

Raza Ur Rahman^{1,2}, Iftikhar Ahmad⁴, Zixiu Li³, Robert P Sparks¹, Amel Ben Saad¹ & Alan C Mullen^{1,2}✉

Single cell RNA sequencing (scRNA-seq) has revolutionized the study of gene expression in individual cell types, but scRNA-seq studies have focused primarily on expression of protein-coding genes. Long noncoding RNAs (lncRNAs) are more diverse than protein-coding genes, yet remain underexplored in part because they are underrepresented in reference annotations applied to scRNA-seq. Merging annotations containing 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, a Singularity image that integrates protein-coding and lncRNA gene transfer format (GTF) annotations to generate enhanced annotations that take into account the sense and antisense overlap of annotated genes, maps scRNA-seq data, and produces files for downstream analysis and visualization. With Singletrome, we detected thousands of lncRNAs not included in GENCODE, clustered cell types based solely on lncRNA expression, and demonstrated that machine learning can predict cell type and disease through lncRNAs alone. 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 with the ability to integrate new lncRNA annotations as they become available.

Long noncoding RNAs (lncRNAs) comprise a diverse class of transcripts that regulate pathology, including cancer¹, immunity², and liver disease³. lncRNA transcripts are at least 200 nucleotides in length, 5' capped, 3' polyadenylated, and are not known to code for proteins⁴. The functions of individual lncRNAs are diverse, with new activities described as additional lncRNAs are investigated. For example, *HOTAIR* promotes tumor invasion and metastasis through interaction with Polycomb repressive complex 2⁵. *HOXA11os* regulates myeloid cell function, binding to a protein in the electron transport chain to maintain a non-inflammatory phenotype⁶, and *DIGIT* (*GSC-DT*) interacts with BRD3 to control definitive endoderm differentiation⁷.

Many lncRNAs also exhibit cell-type-specific patterns of expression⁸. For example, *LOC646329* is enriched in single radial glia of the human neocortex⁸, and *Lnc18q22.2* is induced only in hepatocytes in the setting of metabolic dysfunction-associated steatohepatitis (MASH/NASH)⁹. 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 genes¹⁰, but databases such as LncExpDB and Noncode now report over 100,000 human lncRNA genes^{11,12}.

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 Ranger¹³ exclude reads mapping to exons from different genes that overlap on the same strand; therefore, expanding the number of annotated lncRNA exons may lead to exclusion of 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. The widely used dUTP protocol for stranded RNA-seq¹⁴ can generate spurious antisense reads ranging from 0.6–3% of the sense signal^{15,16}. Analysis of 199

¹Division of Gastroenterology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA. ²Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. ³Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, USA. ⁴Department of Software Engineering, University of Europe for Applied Sciences, Potsdam, Germany. ✉email: alan.mullen@umassmed.edu

strand-specific RNA-seq datasets discovered that spurious antisense reads are detected in these experiments at levels greater than 1% of sense gene expression levels¹⁷. 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 expression data¹⁸. 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 ~ tenfold lower levels than protein-coding genes^{19,20}. 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.

While lncRNAs have been analyzed within particular sets of transcripts, cell types, and datasets using scRNA-seq data^{8,21}, no systematic efforts have analyzed all annotated lncRNAs in scRNA-seq data. Here we develop Singletrome, a framework to create 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. Singletrome is a Singularity image that takes two GTF annotations as input, one containing protein-coding genes and the other containing lncRNAs to generate an enhanced genome annotation. It enables browser extensible data (BED) file creation for downstream analysis, executes Cell Ranger for scRNA-seq data mapping, and merges multiple samples into a single binary alignment map (BAM) file for quality control analysis (RSeQC) and visualization (BigWig). While we applied Singletrome to merge protein-coding annotations from GENCODE with lncRNA annotations from LncExpDB to analyze scRNA-seq data from human peripheral blood mononuclear cells (PBMCs) and liver, our tool can integrate protein-coding and lncRNA gene annotations from any species that are presented in GTF format.

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) and 10,492 protein-coding genes (47,057 exons) that 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 high-confidence antisense lncRNAs from potential artifacts, we developed a trimmed lncRNA genome annotation (TLGA) to retain all the non-overlapping lncRNA exonic regions within the remaining lncRNAs (Fig. 1C). More specifically, we removed the regions of lncRNA exons that coincided with protein-coding exons on the opposite strand with a buffer that included an additional 100 nucleotides (nt) in each direction. We retained lncRNA exons that were a minimum of 200 nt in length after this trimming process. 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. Applying this strategy, we retained 11,673 of the 14,212 lncRNA genes that contain regions antisense to protein-coding genes.

The final TLGA annotation contained 110,599 genes consisting of 19,384 protein-coding genes from GENCODE and 91,215 lncRNA genes from LncExpDB. TLGA increased the number of annotated 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.

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, antisense lncRNAs often have functional activity^{22,23}, 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 still removed 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 (GSE115469²⁴) and liver set 2 (GSE136103²⁵) scRNA-seq data with ULGA and TLGA to assess the output from each annotation. ULGA captures more lncRNA reads than TLGA by retaining exons that overlap protein-coding genes in the antisense direction (i.e., no trimming). In contrast, protein-coding gene Unique Molecular Identifier (UMI) counts remain largely consistent between TLGA and ULGA, but both show a modest reduction compared to GENCODE (Supplementary Table 1). This reflects improved read-mapping specificity: in both TLGA and ULGA, a fraction of reads previously assigned uniquely to protein-coding exons in GENCODE are now also mapped to newly-annotated lncRNA exons and excluded due to multimapping, resulting in fewer reads assigned to protein-coding genes (Fig. 1D). Furthermore, lncRNAs are expressed at

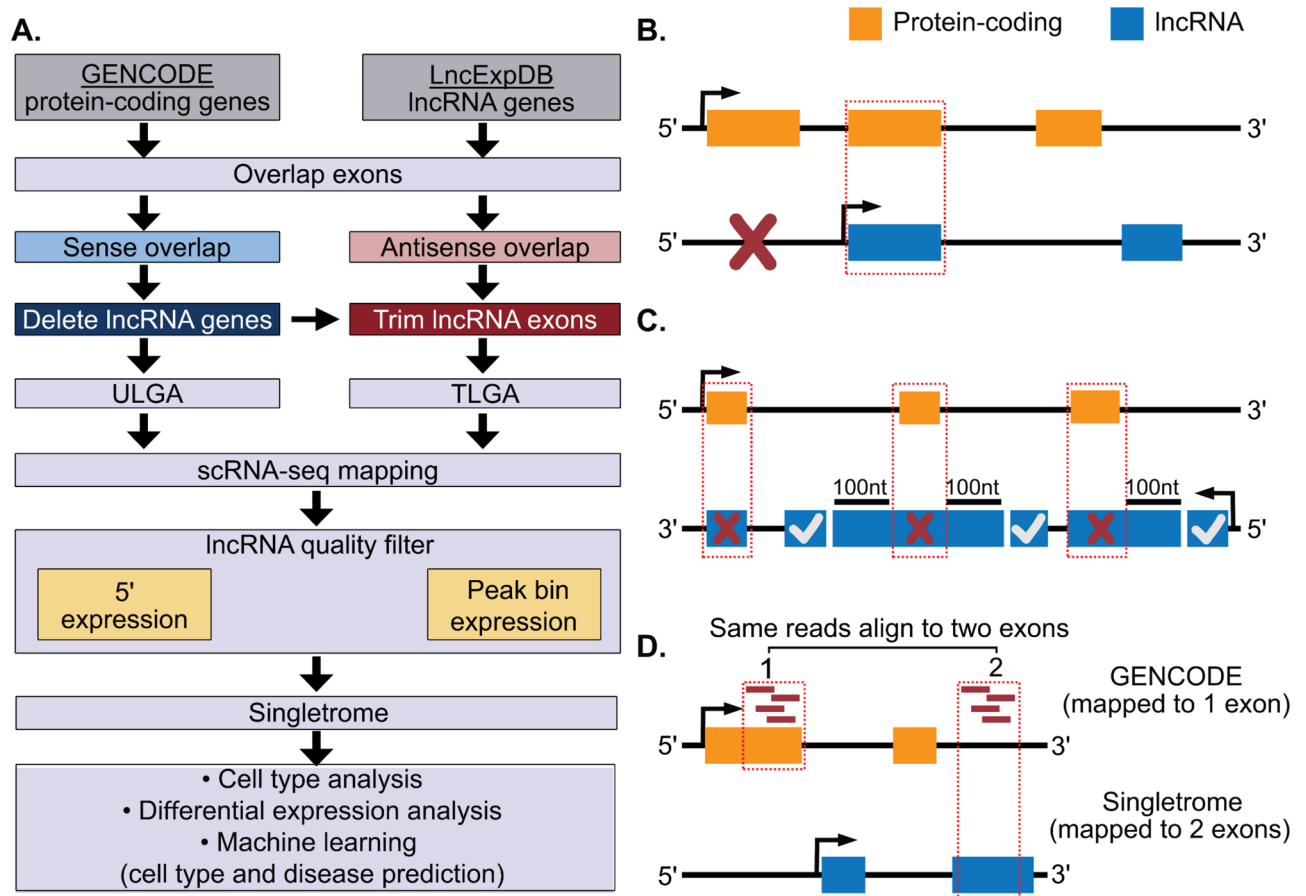


Fig. 1. Enhancing the transcriptome with expanded lncRNA annotation for single cell analysis. **(A)** Development of Singletrome workflow. Exons of protein-coding genes from GENCODE v32 and lncRNA genes from LncExpDB v2 were integrated (Table 1). lncRNA genes overlapping on the sense strand with protein-coding 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 assigned 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 100 nt 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 200 nt (marked with white check). Gene and transcript 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 (a protein-coding exon and an lncRNA exon; red bars 1 and 2). Removing these reads from UMI counting improves read mapping specificity but could reduce the total number of uniquely mapped reads.

lower levels than protein-coding genes across all three datasets (Fig. 2A-B and Supplementary Fig. 1A-D), consistent with previous reports^{19,20}.

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 2). Reads mapped to these lncRNA genes were aligned only to regions antisense to

Feature	Genes		Exons	
	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 v32¹⁰ and LncExpDB v2¹¹. 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). This table summarizes the filtering of lncRNAs based on overlaps with protein-coding genes on the sense and antisense strands, as detailed in the text. A total of 7,531 lncRNAs overlapping protein-coding genes on the sense strand were removed entirely. Of the 14,212 lncRNAs overlapping on the antisense strand, overlapping exon regions were trimmed using a 100 nt buffer; subsequently, 2,539 lncRNAs with exons shorter than 200 nt after trimming were discarded. After these filtering steps, the final set of lncRNAs retained in the Singletrome annotation consists of 91,215 genes (101,285 total – 7,531 sense overlaps removed – 2,539 antisense-trimmed lncRNAs removed). Protein-coding genes and exons were not subjected to filtering (* denotes no filtering applied).

Feature	Genes		Exons	
	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.

protein-coding exons (+ 100 nt) in ULGA and have the potential to come only from library preparation artifacts. These lncRNAs were removed from downstream ULGA 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 reads that could support expression of lncRNA genes where there was corroborating evidence for expression from reads that were not antisense to other exons. 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 2). 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, 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 *SRGAP2C* (blue arrows). *SRGAP2-AS1* is defined as not expressed in TLGA and is excluded from further analysis. *HSALNG0137471* is expressed antisense to *DDX3X* (Fig. 2F). In this example, there are reads supporting expression of *HSALNG0137471* in regions of exons that are not antisense to *DDX3X* exons (black arrows). This lncRNA is defined as expressed in TLGA. There are additional reads mapped in ULGA near 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). We observed an increase in genes detected in both ULGA and TLGA compared to GENCODE (Supplementary Fig. 3A, Supplementary data 1). 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 and 24,034, 19,692, and 38,576 in TLGA (Supplementary material, Supplementary Table 3). Across the 26 samples, ULGA showed a 1.46% increase (118.09 million reads) and TLGA a 1.33% increase (107.41 million reads) in uniquely mapped exonic reads compared to GENCODE (Supplementary Fig. 3B and Supplementary data 1). These results suggest that a fraction of reads not mapped to GENCODE exons are now uniquely mapped to lncRNAs, increasing the number of lncRNA genes detected. In contrast, we observed a decrease in the number of reads uniquely mapped to intronic regions, 0.69% (56 million reads) in ULGA and 0.65% (52.5 million reads) in TLGA compared to GENCODE. Similarly, uniquely mapped intergenic reads also decreased by 1.98% (159.93

million reads) in ULGA and 1.88% (152.23 million reads) in TLGA (Supplementary Fig. 4A, 4B). The reduced percentage of intronic reads in the GENCODE annotation suggest that a small fraction of reads defined as intronic in GENCODE are now annotated as lncRNA exons, while a larger fraction of reads defined as intergenic in GENCODE are now annotated as introns or exons for lncRNAs.

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 anomalies²⁶ (Material and methods). lncRNA genes for which all transcripts met either criterion in a dataset were excluded from further analysis in that dataset.

To obtain the read distribution across the transcript body, we utilized RSeQC²⁷. RSeQC scales all the transcripts to 100 bins, 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 read distribution across lncRNA and protein-coding transcripts was enriched towards the 3' end in PBMCs, as expected for traditional 3' based scRNAs-seq (Fig. 3A), while a larger fraction of reads were located towards the 5' end of lncRNA transcripts in liver set 1 and liver set 2 compared to protein-coding transcripts (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 4). We observed that lncRNA transcripts from 1000–10,000 nt in length contain relatively more 5' reads than protein-coding genes of matched lengths (Fig. 3B–C and Supplementary Fig. 6–8).

We next assessed the variability between gene and transcript lengths and found lower correlation between gene and transcript length for lncRNAs compared to protein-coding genes (Fig. 3D–F & Supplementary Table 5). This finding suggested that the 5' enrichment observed in bulk analysis of lncRNA transcripts might be explained in part 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 1000 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 6). 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 these 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 (Fig. 3, Supplementary Table 6).

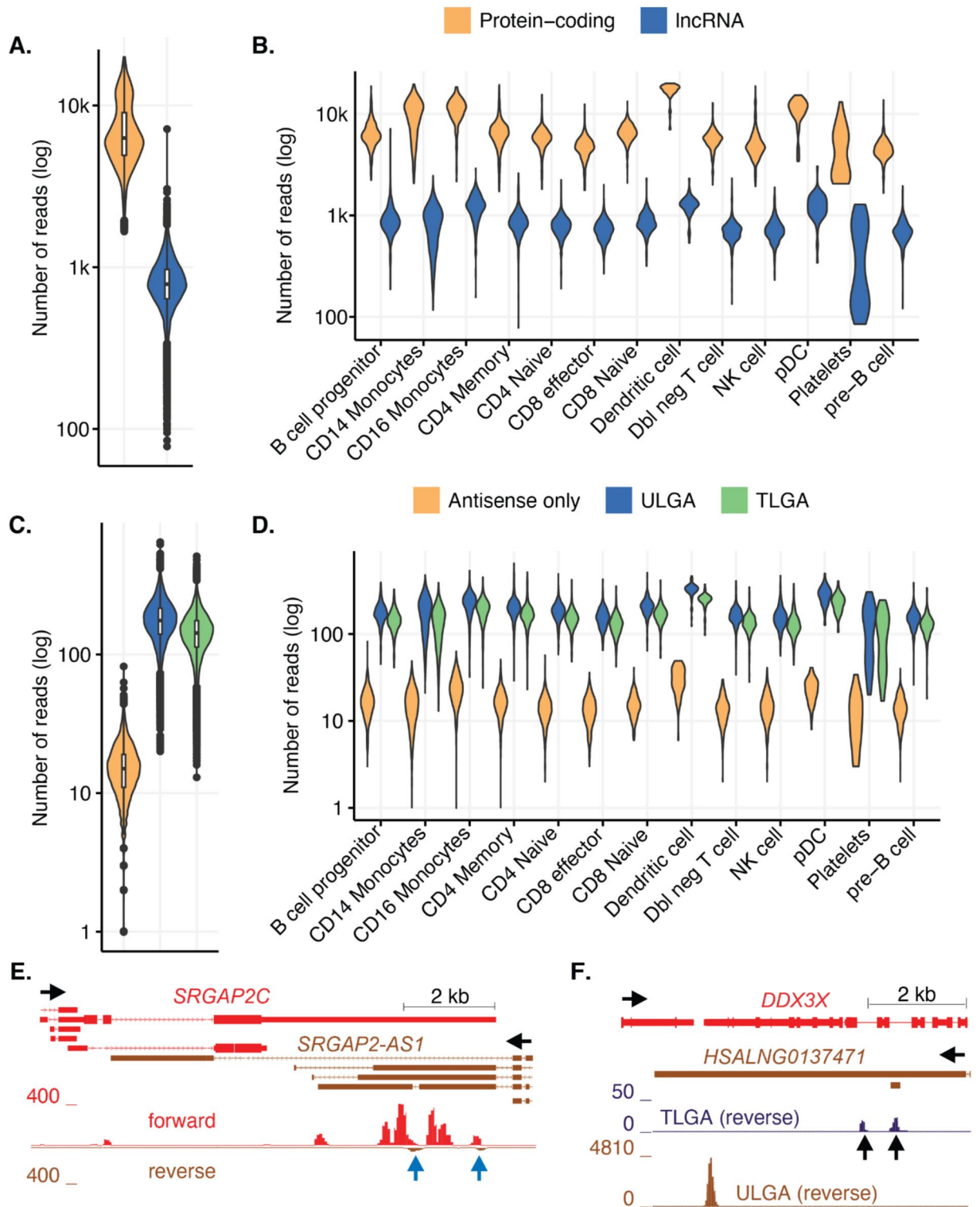
Finally, we evaluated read distribution across lncRNA transcripts to identify potential library artifacts or mapping anomalies. We flagged lncRNA transcripts where reads aligned to one region (minimum transcript length 1000 nt). If the expression of a single bin was greater than the expression of the 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 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 6).

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, respectively (Supplementary Table 6). These lncRNAs were used for downstream analysis.

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

lncRNA expression can be cell-type-specific⁸, and we applied our new annotation to determine if we could group 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. In current practice, clustering is based primarily on protein-coding genes, and lncRNAs alone are not typically used for clustering. Therefore, as an initial step, we did not perform clustering. Instead, we used a standard Seurat workflow to generate UMAP projections based on principal component analysis (PCA) of variable genes, allowing us to visualize the data. We then examined whether the original cell type labels retained separation in UMAP space when the embedding was based solely on lncRNA expression. We generated UMAP projections of cells using data aligned to both GENCODE and Singletrome, applying the previously established filters. Despite lower expression of lncRNAs compared to protein-coding genes (Fig. 2A–B, Supplementary Fig. 1A–D), the UMAP projections based on lncRNAs alone revealed clear grouping of many cell types for both PBMCs (Fig. 4A–D) and liver (Supplementary Fig. 9A–D, Supplementary Fig. 10A–D).

To assess the added value of lncRNA annotations in Singletrome compared to lncRNA annotations in GENCODE, we next performed clustering using only lncRNAs from each annotation source (Material and methods). In the PBMC dataset, clustering with Singletrome lncRNAs resulted in 16 clusters, compared to 10 clusters using GENCODE lncRNAs (Supplementary Fig. 11A–D, 12A–D). This also yielded a higher Adjusted Rand Index (ARI) (5.5% increase) when comparing Singletrome lncRNA and protein-coding gene clustering (0.500) to GENCODE lncRNA and protein-coding gene clustering ARI (0.445) (Material and methods). These results suggest that Singletrome lncRNAs clusters align more consistently with clusters from protein-coding genes and offer higher resolution. For example, when comparing clustering with GENCODE lncRNAs to the original cell type labels, 458 of 460 B-cell progenitors and 958 of 959 pre-B cells are grouped into a single cluster.



In contrast, clustering with Singletrome lncRNAs separates them into two distinct clusters: cluster 4, containing 879 pre-B cells and 43 B-cell progenitors, along with one NK cell and one CD16+ monocyte; and cluster 9, containing 417 B-cell progenitors, 80 pre-B cells, and one pDC. Alluvial plots show the redistribution of cells from 20 protein-coding based clusters into 10 clusters when using GENCODE lncRNAs and into 16 clusters when using Singletrome lncRNAs (Supplementary Fig. 13A-B).

Analysis of liver datasets showed similar results. In liver set 1 (GSE115469), clustering with Singletrome lncRNAs produced 16 clusters, and GENCODE lncRNAs resulted in 13 clusters. A higher ARI was again observed for Singletrome lncRNAs (0.342 compared to 0.338, ~0.4% improvement) (Supplementary Fig. 14A-D, 15A-D). Alluvial plots further illustrate the redistribution of cells between cluster assignments based on Singletrome and GENCODE lncRNAs (Supplementary Fig. 16A-B). In liver set 2 (GSE136103), Singletrome lncRNAs produced 31 clusters and GENCODE lncRNAs yielded 51 clusters. The ARI was again higher for Singletrome lncRNAs

Fig. 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 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 downstream analysis after the lncRNA is defined as expressed based on TLGA reads.

Dataset	Source	Number of samples	Number of cells
pbmc_10k_v3 (10x Genomics)*	PBMCs	1	9432
GSE115469 ²⁴	Liver	5	8444
GSE136103 ²⁵	Liver	20	58,358

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

compared to GENCODE lncRNAs (0.338 compared to 0.252, 8.5% improvement). These results highlight how the enhanced lncRNA annotation in Singletrome improved the ability to capture distinct cell groupings (Supplementary Fig. 17A–D and 18A–D). Alluvial plots (Supplementary Fig. 19A–B) further illustrate the shift in cluster assignments between Singletrome and GENCODE lncRNA-based clustering. Since lncRNAs can cluster the majority of cells by cell type, we next aimed to generate an lncRNA-based cell type marker map using cell labels from the original publications. 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. 20–23). While lncRNAs are expressed at lower levels compared to protein-coding genes in all datasets (Fig. 4G, Supplementary Fig. 24), 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)²⁸ 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. 25–26 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. 27 Supplementary data 7). Liver set 1 separates single cell type into multiple clusters based on marker genes from GENCODE. For example, there are six cell clusters of hepatocytes, three clusters of liver sinusoidal endothelial cells (LSECs), and two cell clusters of each macrophages and gd T cells. To assess the accuracy of predicting cell types rather than sub-clusters of cell types for liver set 1, we merged clusters within the same cell type, retaining 11 cell types. 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. 28 and Supplementary data 7). These results serve as proof-of-concept that lncRNA expression can predict cell types with accuracy comparable to protein-coding genes, even though we

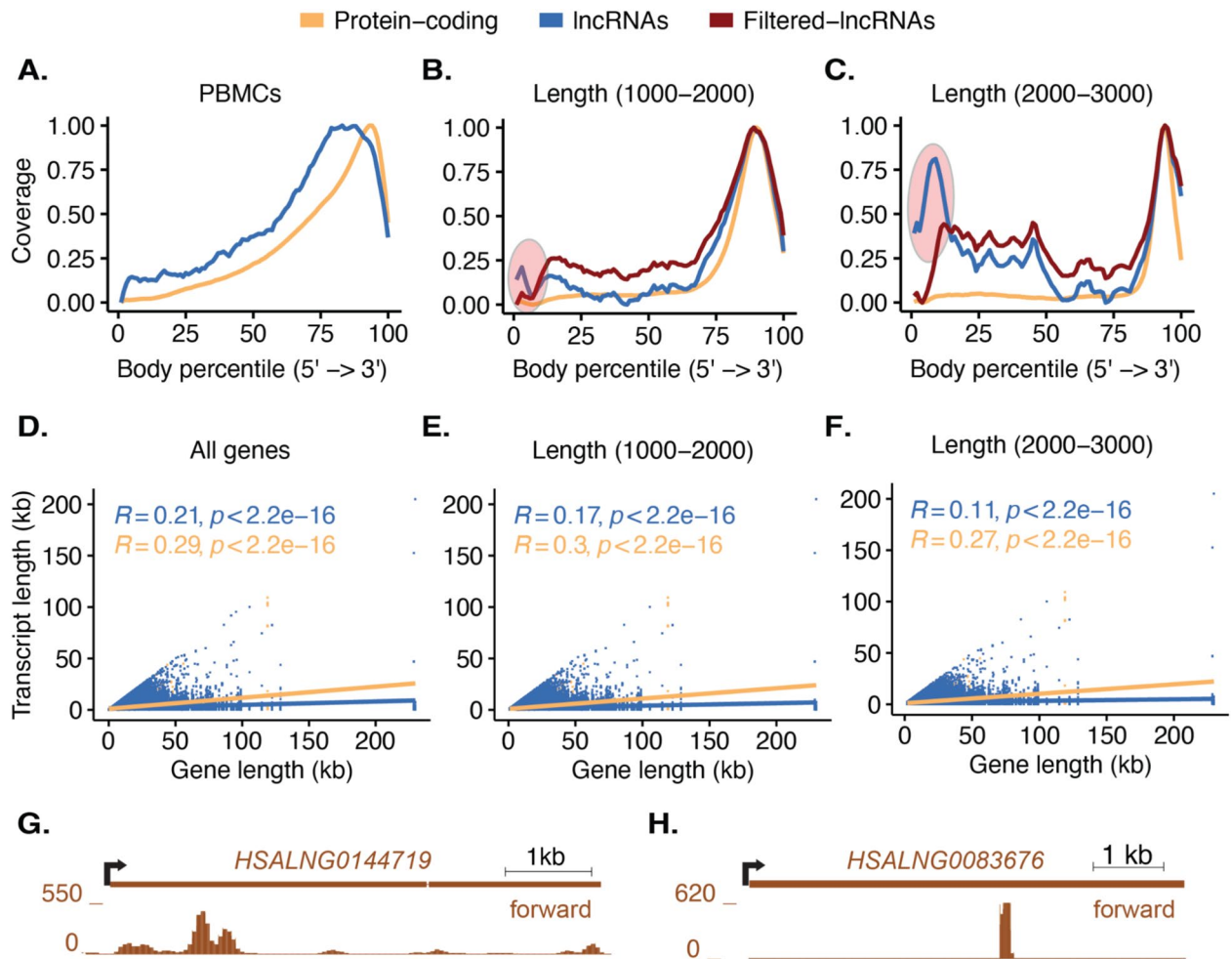


Fig. 3. Distribution and quality control of lncRNA mapping in PBMCs. **(A)** Distribution of reads mapped across transcripts of protein-coding genes (orange) and lncRNA genes (blue). The x-axis represents RNA transcripts from 5' to 3' divided into 100 bins (Body percentile), and the y-axis indicates 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 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 mapped reads after removing lncRNAs that were flagged for low quality. **(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 correlation coefficient (R), for lncRNAs is shown in blue and the value for protein-coding genes is shown 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.

anticipate that including both lncRNAs and protein-coding genes would be the primary approach to define cell types.

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, GSE136103²⁵). We again used cell labels from the original study and visualized the cells with Uniform Manifold Approximation and Projection (UMAP) based on the condition (healthy and cirrhotic) using Singletrome (Supplementary Fig. 29), protein-coding genes from Singletrome (Supplementary Fig. 30), and lncRNAs from Singletrome (Fig. 5A). We performed differential expression analysis of lncRNAs in healthy and cirrhotic liver by cell type.

We detected 937 differentially expressed lncRNA genes (495 upregulated and 442 downregulated) 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 ($\text{padj} < 0.1$ and $\log_2\text{FC} > 0.25$, 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 *H19* and *MEG3* in mesenchymal cells (Fig. 5C, Supplementary data 8). *H19* has been shown to promote liver fibrosis^{29,30} while *MEG3* has been linked to both pro- and anti-fibrotic phenotypes in the liver^{31,32}. While *XIST* was also enriched in specific cell types in cirrhotic samples (Fig. 5C), this gene is located in the X-chromosome, and expression levels will be affected by the distribution of sexes within this dataset (cirrhotic: 2 female and 3 male; healthy: 1 female and 4 male) (Material and methods). Additional lncRNAs, including *HSALNG0146932* and *HSALNG0035811* in mesenchymal cells, *HSALNG0061639* and *HSALNG0016726* in cholangiocytes, and *HSALNG0142578* in hepatocytes were also found to be enriched in cirrhotic livers and are not annotated in GENCODE (Fig. 5C–F, Supplementary data 8). 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 group cell types and subtypes (Fig. 4D, 4H and Supplementary Fig. 9D, 10D), and we were able to observe differences in lncRNA expression between healthy and cirrhotic livers in individual cell types. We next trained a machine learner (XGBoost) on Singletrome lncRNA from expression data from liver set 2 (GSE136103) and predicted the condition (healthy or cirrhotic) for the target cells. 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 model on the expression data of protein-coding genes, and we were able to predict the condition of 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.

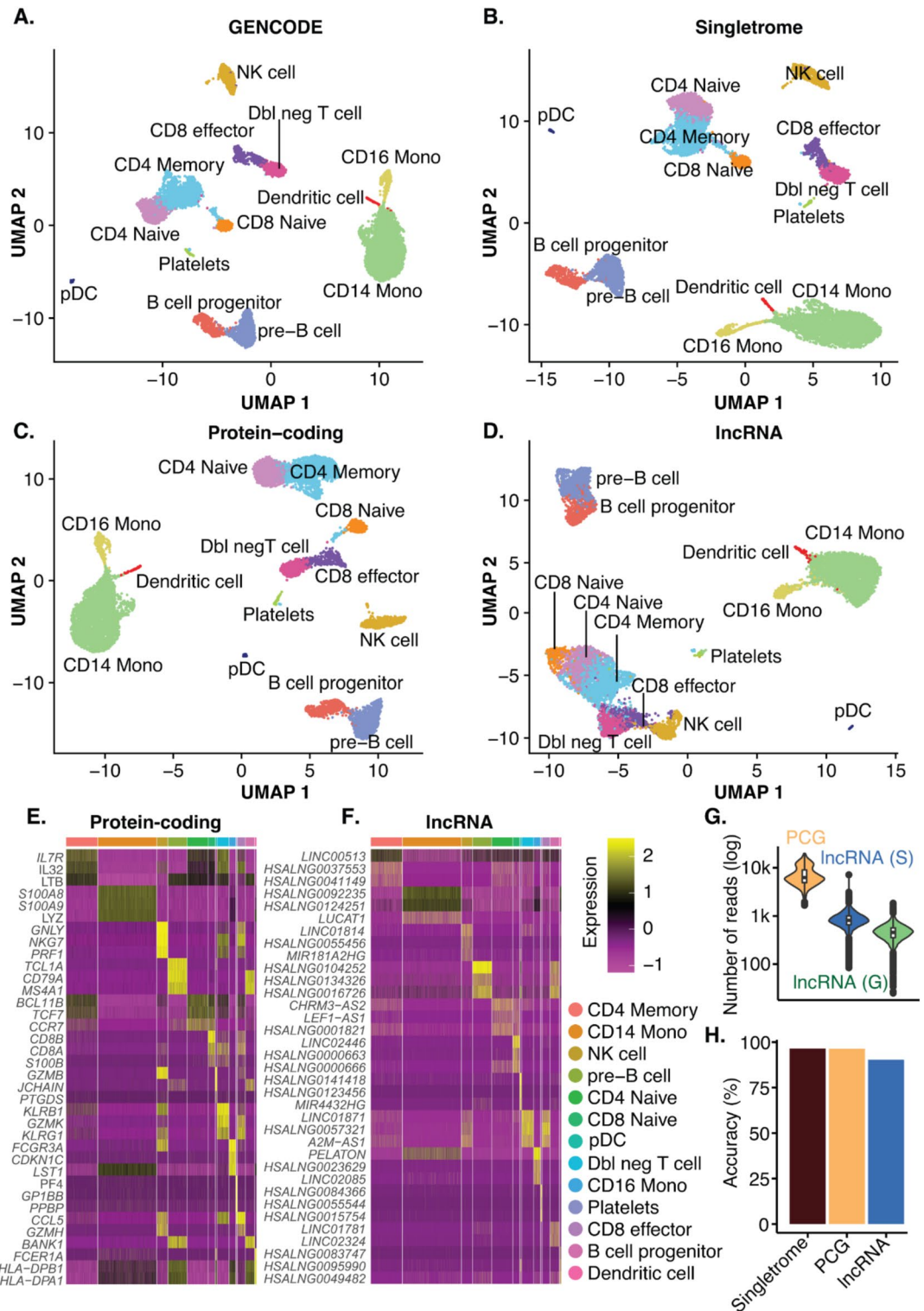
Discussion

Analysis of lncRNAs has been performed in scRNA-seq data for particular sets of transcripts and cell types^{8,21}, but more universal lncRNA pipelines are not available. Here, we develop Singletrome, a unified analysis framework that can quantify lncRNA expression in any human scRNA-seq data with the goal of increasing the depth of single cell annotations to define differentially expressed lncRNA genes that may regulate cell function in health and disease.

Singletrome is a Singularity image that integrates two GTF annotations, one containing protein-coding genes and another containing lncRNAs to generate an enhanced genome annotation. This tool provides a streamlined workflow for downstream analyses by creating BED files for further processing, running Cell Ranger for scRNA-seq data mapping, and merging multiple samples into a single BAM file for RSeQC quality control analysis. Additionally, it produces BigWig files for visualization. By accepting and merging two GTF files, Singletrome can accommodate a variety of annotation sources, making it a versatile tool for researchers working with different genome annotations. Our application of Singletrome focused on the LncExpDB database, but the tool is equally compatible with other lncRNA annotation sources. It can also be extended to any organism where both lncRNAs and protein-coding genes are defined in GTF format, providing broad applicability for genomic studies across various species.

In applying Singletrome, we removed lncRNAs that shared exons with protein-coding genes, and we utilized trimmed lncRNA genome annotation (TLGA) to avoid counting spurious antisense reads when defining lncRNAs that are expressed in a dataset. We then applied 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 2). For downstream analysis, we next removed lncRNAs where mapped reads exhibit 5' bias in 3' sequenced scRNA-seq libraries and where the majority of reads were mapped to a single region within the transcript, as both situations could represent library artifacts or mapping anomalies²⁶ (Fig. 3G–H and Supplementary Table 6). While this approach reduces the total number of reads that can be mapped to lncRNAs and decreases the number of lncRNAs that can be identified, it increases our confidence in those lncRNAs determined to be expressed. Applying this approach to three publicly available 10x scRNA-seq datasets demonstrated an increase in the detection of lncRNA genes by greater than five fold compared to the use of GENCODE, while providing further confidence in the data supporting expression of these lncRNAs.

lncRNA expression can be cell-type-specific⁸, and we found that most cell types can be clustered by lncRNAs alone (Supplementary Fig. 11, 12, 14, 15, 17 and 18). In addition, the expanded lncRNA annotations provided by Singletrome further improved clustering compared to lncRNAs in GENCODE, as quantified by



ARI (Supplementary data 9). These findings show that Singletrome expands the identification of lncRNAs differentially expressed across cell types. Comparing cells from healthy and cirrhotic liver (liver set 2), we were then able to identify 937 differentially-expressed lncRNAs. *H19* and *MEG3* are included in GENCODE, and were identified in our analysis. Both lncRNAs modulate liver fibrosis^{30–32}, suggesting that additional lncRNAs identified with similar patterns of expression (Fig. 5C–F and Supplementary data 8) may also have activity in liver fibrosis/cirrhosis. This analysis was based on available data for healthy and cirrhotic liver. As datasets expand in the future to include additional replicates of healthy and diseased tissue, greater statistical power will facilitate the identification of differentially-expressed lncRNAs across many different diseases.

While the application of machine learning to scRNA-seq analysis will require expression data from protein-coding genes for cell type identification and may benefit from the additional reads provided by lncRNA annotation, we also wanted to evaluate the concept that cell types can be identified based on the cell type diversity

◀ **Fig. 4.** lncRNAs alone predict most clusters and cell types in single cell data. scRNA-seq data from PBMCs were mapped and visualized with Uniform Manifold Approximation and Projection (UMAP) 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 PBMCs. (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.

of lncRNA expression. To determine cell types based on lncRNAs without the statistical assumptions, we applied the XGBoost classifier, as it is a preferred machine learning technique for classification with imbalanced datasets (high variability in cell numbers in different cell types) often observed in scRNA-seq^{33–35}. The accuracy for predicting cell types using lncRNAs was nearly 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). The differences in accuracy most likely reflects the greater number of transcripts mapped to protein-coding genes, but an inherent bias towards the protein-coding genes in the original cell type labeling may also impact this comparison. Applying machine learning to healthy and cirrhotic liver also demonstrated that lncRNA expression is altered in disease at the level of cell types. These results highlight the depth and diversity of lncRNA transcripts within cell types and subtypes and provide further support to suggest that lncRNAs, in addition to protein-coding genes, can serve as biomarkers and mechanistic drivers of disease^{36–38}.

As of December 2024, the Human Cell Atlas (HCA) has mapped over 62.7 million individual cells from 9,200 donors, spanning 478 projects across a wide range of tissues, bringing us closer to a comprehensive cellular map of the human body^{39–41}. 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.

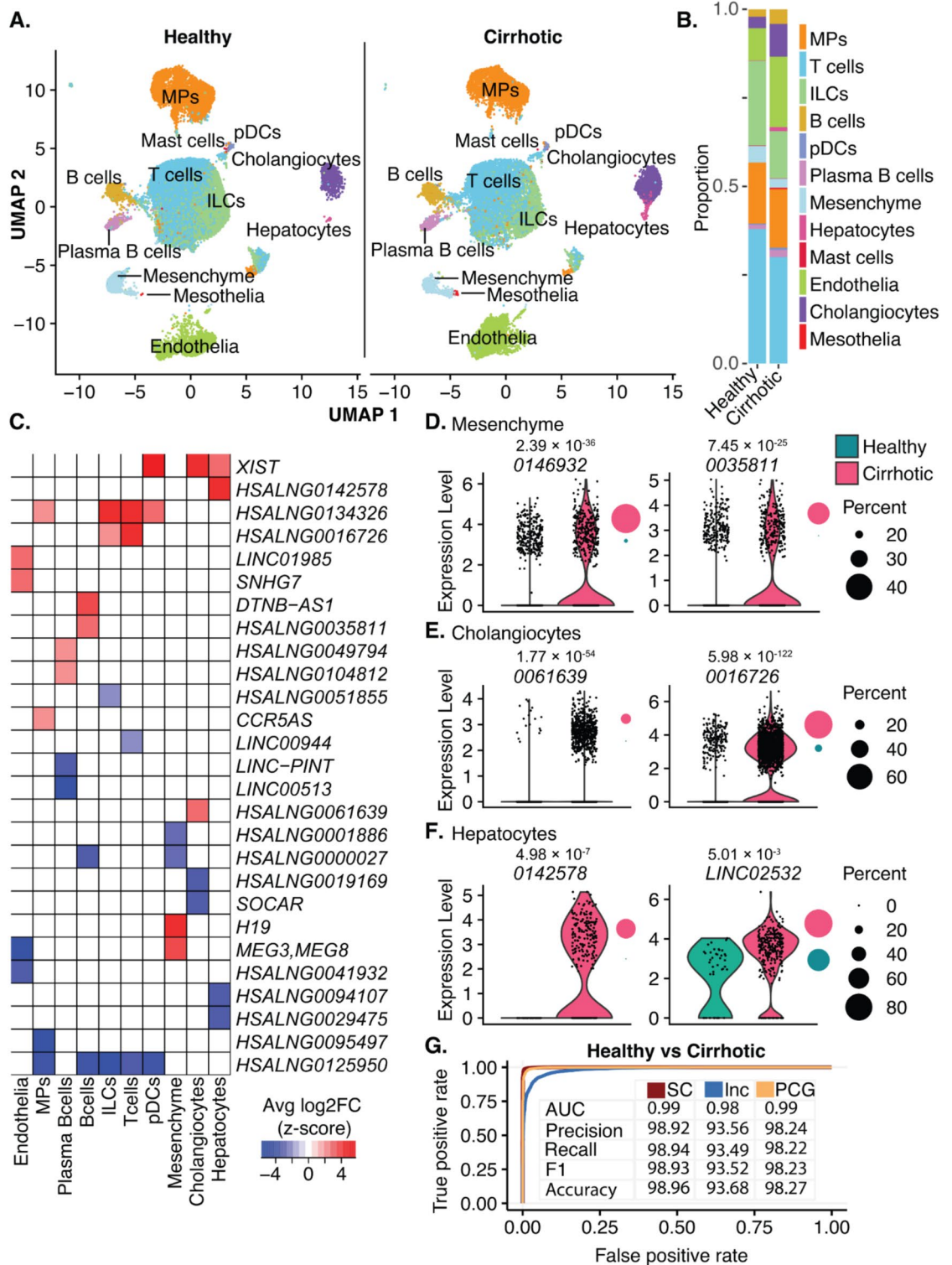
Material and methods

Container environment and dependencies

Singletrome is provided as a Singularity image, which requires an apptainer. Users may download the prebuilt SIF container or rebuild it using the provided `build.apptainer` script (section code availability). The container includes essential bioinformatics tools installed via Miniconda (e.g., BEDTools, RSeQC, DeepTools, and Cell Ranger). The Singletrome pipeline, executed via the `Singletrome.py` script, integrates two GTF annotations, one containing protein-coding genes and another containing lncRNAs. It automates tasks such as downloading and preprocessing GTF files, analyzing exon overlaps using BEDTools, merging annotations into a final GTF file, and builds a genome index for Cell Ranger. Furthermore, if multiple samples are analyzed and mapped using the `run_cellranger.py` script, it will generate a merged BAM file for RSeQC quality control analysis and a BigWig file for visualization.

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 7 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/lncxpd/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 to 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 8). 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 10× 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 default parameters for all genome versions to obtain gene expression count matrix.

◀ **Fig. 5.** lncRNA expression predicts disease pathology. **(A)** scRNA-seq data from liver set 2 (GSE136103) were mapped and visualized with Uniform Manifold Approximation and Projection (UMAP) 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 conditions, 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’ were 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 downregulated) lncRNA genes (y-axis) between healthy and cirrhotic liver for each cell type. Average log₂-fold change 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 (blue) and cirrhotic liver (red). Statistical significance (Materials and methods), based on adjusted p-values (padj), is indicated above gene name. **(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).

lncRNA quality filter

To compute the gene body coverage for each dataset (PBMCs, liver set 1 and liver set 2) we utilized RSeQC²⁷. The program was used to check if read coverage was uniform, if there was any 5′ or 3′ end bias, or if the majority of reads mapped to one location (single bin) in the transcript. RSeQC scales each transcript 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 4). 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 anomalies²⁶. For 5′ bias transcripts detection, we flagged transcripts ≥ 1000 nucleotides in length if 50% or more of the total read coverage was located within the first 10 bins (representing the 5′ end). To identify transcripts with the majority of reads mapped to a single location, we flagged transcripts where a single bin within the first 90 bins accounted for more reads than the combined total of the remaining 99 bins. This filter was also applied only to lncRNAs with transcript lengths ≥ 1000 nucleotides. lncRNA genes for which all transcripts met either criterion 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, adjusted rand index (ARI), and alluvial plot analysis

All single-cell RNA-seq gene expression matrices were analyzed using Seurat (v4.0.6). We retained original cell type labels by matching cell barcodes from our expression matrices to those provided in the original publications (see Table 3). Only cells with a defined label in the original dataset were retained for downstream analyses.

To evaluate the contribution of lncRNA annotations in Singletrome relative to those in GENCODE, we performed clustering separately using only lncRNAs from each annotation source (GENCODE and Singletrome). To compare, protein-coding gene-based clustering was also performed from the same annotation source (GENCODE or Singletrome). For each dataset, we subsetted the Singletrome count matrix into protein-coding and lncRNA gene matrices prior to clustering.

Clustering was performed using Seurat’s standard workflow, including normalization, scaling, PCA, and UMAP dimensionality reduction. We used a clustering resolution of 1.0 to obtain a higher number of clusters and maintained this setting for both protein-coding and lncRNA-based clustering to allow direct comparisons. In all UMAPs, cluster identity was annotated using the original cell type labels from the publications.

To quantify the similarity between clusterings derived from lncRNAs and protein-coding genes, we used the Adjusted Rand Index (ARI)⁴². ARI scores were calculated using the ‘adjustedRandIndex()’ function in the mclust R package (v6.1.1). This function computes an index that quantifies agreement between two cluster assignments. ARI values range from 0 (random agreement) to 1 (perfect concordance). For each dataset, ARI was computed by comparing the lncRNA-derived clusters to the protein-coding gene-derived clusters, treating the latter as ground truth. This was done separately for GENCODE and Singletrome annotations. The difference in the ARI between Singletrome and GENCODE lncRNA annotations was used as a comparative measure.

To visualize how cells transitioned between protein-coding and lncRNA-based clusters, we generated Alluvial plots for each dataset and annotation type. These plots illustrate the flow of cells from clusters defined by protein-coding genes to those defined by lncRNAs. Cluster labels were assigned based on the majority cell type identity from the original dataset within each cluster. Detailed cluster composition and label assignment are provided in Supplementary data 9.

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 underrepresented 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 the FindMarkers function from Seurat, which applies a Wilcoxon Rank Sum test by default, to identify 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 or equal to 0.1 and `log2FC` more than 0.25 in either direction (Supplementary data 8).

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).

Sex determination of liver set 2 samples

The original publication²⁵ and GEO dataset (GSE136103) did not provide sex information for the Liver Set 2 samples. To infer the sex of each sample, we analyzed the expression of X- and Y-chromosome-associated genes, including *RPS4Y1*, *XIST*, and *ZFX*. High expression of Y-linked genes (*RPS4Y1*) was used as an indicator of male samples, whereas the presence of *XIST*, an X-chromosome gene involved in X-inactivation, suggested female samples. Based on this approach, we inferred that the cirrhotic group consisted of two female and three male samples, while the healthy group included one female and four male samples.

Data availability

Data availability. All the datasets (Table 3) used in this study are publicly available. The PBMC dataset was obtained from the 10x Genomics platform “10 k 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. Code availability. Python, R and Bash Scripts for data processing are available through <https://github.com/RAZA-UR-RAHMAN/Singletrome>.

Code availability

Python, R and Bash Scripts for data processing are available through <https://github.com/RAZA-UR-RAHMAN/Singletrome>.

Received: 3 March 2025; Accepted: 24 July 2025

Published online: 12 August 2025

References

- Huang, W. et al. The long non-coding RNA SNHG3 functions as a competing endogenous RNA to promote malignant development of colorectal cancer. *Oncol. Rep.* **38**, 1402–1410 (2017).
- Kotzin, J. J. et al. The long non-coding RNA Morrbid regulates Bim and short-lived myeloid cell lifespan. *Nature* **537**, 239–243 (2016).

3. Mahpour, A. & Mullen, A. C. Our emerging understanding of the roles of long non-coding RNAs in normal liver function, disease, and malignancy. *JHEP Rep.* **3**, 100177 (2021).
4. Staleno, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell. Biol.* **22**, 96–118 (2021).
5. Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
6. Shmuel-Galia, L. et al. The lncRNA HOXA11os regulates mitochondrial function in myeloid cells to maintain intestinal homeostasis. *Cell Metab.* **35**, 1441–1456.e9 (2023).
7. Daneshvar, K. et al. lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation. *Nat. Cell. Biol.* **22**, 1211–1222 (2020).
8. Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.* **17**, 67 (2016).
9. Atanasovska, B. et al. A liver-specific long noncoding RNA with a role in cell viability is elevated in human nonalcoholic steatohepatitis. *Hepatology* **66**, 794–808 (2017).
10. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
11. Li, Z. et al. LncExpDB: an expression database of human long non-coding RNAs. *Nucleic Acids Res.* **49**, D962–D968 (2021).
12. Fang, S. et al. NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **46**, D308–D314 (2018).
13. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
14. Parkhomchuk, D. et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
15. Zeng, W. & Mortazavi, A. Technical considerations for functional sequencing assays. *Nat. Immunol.* **13**, 802–807 (2012).
16. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
17. Mourão, K. et al. Detection and mitigation of spurious antisense expression with RoSA. *F1000Res* **8**, 819 (2019).
18. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
19. Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
20. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
21. Luo, H. et al. Single-cell Long Non-coding RNA Landscape of T Cells in Human Cancer Immunity. *Genom. Proteom. Bioinform.* **19**, 377–393 (2021).
22. Faghihi, M. A. et al. Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol.* **11**, R56 (2010).
23. Yap, K. L. et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell.* **38**, 662–674 (2010).
24. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 1–21 (2018).
25. Ramachandran, P. et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512–518 (2019).
26. Ma, C. & Kingsford, C. Detecting, Categorizing, and Correcting Coverage Anomalies of RNA-Seq Quantification. *Cell Syst.* **9**, 589–599.e7 (2019).
27. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM. <https://doi.org/10.1145/2939672.2939785> (2016).
29. Wu, X.-J., Xie, Y., Gu, X.-X., Zhu, H.-Y. & Huang, L.-X. LncRNA XIST promotes mitochondrial dysfunction of hepatocytes to aggravate hepatic fibrogenesis via miR-539-3p/ADAMTS5 axis. *Mol. Cell Biochem.* **478**, 291–303 (2022).
30. Xiao, Y. et al. Long Noncoding RNA H19 Contributes to Cholangiocyte Proliferation and Cholestatic Liver Fibrosis in Biliary Atresia. *Hepatology* **70**, 1658–1673 (2019).
31. Zhang, L., Yang, Z., Trottier, J., Barbier, O. & Wang, L. Long noncoding RNA MEG3 induces cholestatic liver injury by interaction with PTBP1 to facilitate shp mRNA decay. *Hepatology* **65**, 604–615 (2017).
32. Yu, F., Geng, W., Dong, P., Huang, Z. & Zheng, J. LncRNA-MEG3 inhibits activation of hepatic stellate cells through SMO protein and miR-212. *Cell Death Dis.* **9**, 1014 (2018).
33. Hernesniemi, J. A. et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. *Ann. Med.* **51**, 156–163 (2019).
34. Nishio, M. et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS ONE* **13**, e0195875 (2018).
35. Ogunleye, A. & Wang, Q.-G. XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 2131–2140 (2020).
36. Nath, A. et al. Discovering long noncoding RNA predictors of anticancer drug sensitivity beyond protein-coding genes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 22020–22029 (2019).
37. Delás, M. J. & Hannon, G. J. lncRNAs in development and disease: from functions to mechanisms. *Open Biol.* **7**, 170121 (2017).
38. Bolha, L., Ravnik-Glavač, M. & Glavač, D. Long Noncoding RNAs as Biomarkers in Cancer. *Dis. Markers.* **2017**, 7243968 (2017).
39. Suo, C. et al. Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).
40. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
41. Parums, D. V. Editorial: The Human Cell Atlas. What is it and where could it take us?. *Med. Sci. Monit.* **30**, e947707 (2025).
42. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).

Acknowledgements

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. This publication is part of the Human Cell Atlas—www.humancellatlas.org/publications.

Author contributions

R.R. and A.C.M. conceived and designed the study. Computational analyses were performed by R.R. Z.L. developed and tested the singularity image with input from 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.B.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.

Funding

Chan Zuckerberg Initiative, Pew Biomedical Scholars Program, NIH/NIDDK R01DK116999.

Declarations

Competing interests

A.C.M. has received research funding from Boehringer Ingelheim and GlaxoSmithKline for unrelated projects. R.R. is a co-founder of deepnostiX, based in Germany and Pakistan, and founder of VitalEdge in the USA. Additionally, R. R. serves as a consultant for Ibis Therapeutics and Vigil Neuro. No other authors have conflicts to declare.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-13528-9>.

Correspondence and requests for materials should be addressed to A.C.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025