nature communications

Article

https://doi.org/10.1038/s41467-024-54005-7

Uncovering functional lncRNAs by scRNA-seq with ELATUS

Received: 21 February 2024

Accepted: 28 October 2024

Published online: 09 November 2024

Check for updates

Enrique Goñi[1,2,3], Aina Maria Mas[1,2,3], Jovanna Gonzalez[1,2,3], Amaya Abad[1,2], Marta Santisteban[2,3,4], Puri Fortes[1,2,3,5], Maite Huarte[1,2,3] ✉ & Mikel Hernaez[1,2,3,6] ✉

Long non-coding RNAs (lncRNAs) play fundamental roles in cellular processes and pathologies, regulating gene expression at multiple levels. Despite being highly cell type-specific, their study at single-cell (sc) level is challenging due to their less accurate annotation and low expression compared to protein-coding genes. Here, we systematically benchmark different preprocessing methods and develop a computational framework, named ELATUS, based on the combination of the pseudoaligner Kallisto with selective functional filtering. ELATUS enhances the detection of functional lncRNAs from scRNA-seq data, detecting their expression with higher concordance than standard methods with the ATAC-seq profiles in single-cell multiome data. Interestingly, the better results of ELATUS are due to its advanced performance with an inaccurate reference annotation such as that of lncRNAs. We independently confirm the expression patterns of cell type-specific lncRNAs exclusively detected with ELATUS and unveil biologically important lncRNAs, such as *AL121895.1*, a previously undocumented cis-repressor lncRNA, whose role in breast cancer progression is unnoticed by traditional methodologies. Our results emphasize the necessity for an alternative scRNA-seq workflow tailored to lncRNAs that sheds light on the multifaceted roles of lncRNAs.

Organismal functions are ultimately driven by the orchestration of the transcriptional programs of each of the individual cells that compose their tissues. The profound understanding of the cellular transcriptional configurations allows uncovering the mechanisms underlying pathological processes. While new technologies enable profiling of transcriptomes at single-cell resolution, novel computational methods are crucially needed for exploring transcriptional events at non-coding regions.

Most gene expression studies at single-cell level are exclusively focused on protein-coding genes while non-coding RNA species are very poorly investigated[1–4]. A significant fraction of non-coding RNAs are classified as long non-coding RNAs (lncRNAs), RNA Pol II transcripts lacking protein-coding potential, recently re-defined based on their length of more than 500 nucleotides[5,6]. LncRNAs are distinctively characterized by their high tissue[7] and cell type specificity[8–10] compared to protein-coding genes, which is linked to their regulatory functions. In line with their roles in gene regulation, alterations in lncRNA expression are associated with multiple pathologies[11,12]. All these characteristics evidence the potential benefits from their study at single-cell resolution, which is needed to achieve an improved and complete definition of cellular identity. However, limitations

[1]Center for Applied Medical Research, University of Navarra, PIO XII 55 Ave, Pamplona, Spain. [2]Institute of Health Research of Navarra (IdiSNA), Pamplona, Spain. [3]Cancer Center Clinica Universidad de Navarra (CCUN), Madrid, Spain. [4]Department of Medical Oncology, Breast Cancer Unit, Clinica Universidad de Navarra, Pio XII 36 Ave, Pamplona, Spain. [5]Liver and Digestive Diseases Networking Biomedical Research Centre (CIBERehd), Spanish Network for Advanced Therapies (TERAV ISCIII), Madrid, Spain. [6]Data Science and Artificial Intelligence Institute (DATAI), Universidad de Navarra, Pamplona, Spain. ✉ e-mail: maitehuarte@unav.es; mhernaez@unav.es

such as their low expression and low accuracy of their annotation have greatly hindered their use in this type of studies. According to the conservative and widely used GENCODE annotation, more than 19,000 lncRNAs are present in the human genome[13]. In contrast to the steady quantification of protein-coding genes, the annotation of lncRNAs has been in continuous evolution and growth over the last decade (Supplementary Fig. 1). Besides being less stable, the weak conservation levels of lncRNAs during evolution[14] and their low expression values, complicates their detection in bulk transcriptomic data and makes the mapping of lncRNAs more challenging[15–17], highlighting the need for appropriate computational methods.

Single-cell RNA-sequencing (scRNA-seq) has transformed transcriptomics by enabling the investigation of gene expression in individual cells, providing a comprehensive characterization of tissues[18–20] and allowing the inspection of cell dynamics[21]. Particularly, scRNA-seq droplet-based methods[22,23], predominantly the 10x Genomics technology, have revolutionized the procedure by increasing the throughput of cells and decreasing the sequencing costs[24–26].

The computational pipeline of 10x Genomics scRNA-seq experiments begins with a preprocessing step of the sequenced samples to generate the unfiltered cell-by-gene count matrix[27], which precedes the downstream analysis[28,29]. This critical step involves mapping the reads containing the sequenced cDNAs, as well as correcting both cell and UMI barcodes, in order to identify individual RNA molecules. Different programs based on the aligner STAR[30], that map reads to a reference genome, such as the widely used Cell Ranger[24] (developed by 10x Genomics) or STARsolo[31], perform the entire preprocessing step. In addition, the pseudoaligners Kallisto and Salmon, which are based on matching read k-mers to the transcriptome[32,33], have also incorporated a suite of tools to preprocess the scRNA-seq sequenced reads, named Bustools[34,35] and Alevin[36], respectively. When first released, Kallisto and Salmon were based on similar algorithms, but they have diverged over time due to several modifications, such as the selective alignment strategy incorporated by Salmon[37].

To date, scRNA-seq focused on lncRNAs have been mostly performed using chips or plate-based technologies, such as the Fluidigm C1 microfluidic platform[8,38], or the SMART-seq or SMART-seq2 protocols[2,39–41], respectively. However, these technologies have an important limitation in the number of cells that can be jointly sequenced[26]. More recently, the droplet-based 10x Chromium technology became the dominant protocol for scRNA-seq due to its high yield[42]. Nevertheless, studies investigating lncRNAs via 10x Genomics are still scarce, and have been limited to apply the standard Cell Ranger preprocessing pipeline[1,4,43,44], without testing other options. In this context, previous comparative studies of scRNA-seq preprocessing pipelines[29,31,34,36,45–48], unfortunately, did not focus on the detection and quantification of lncRNAs.

Here, after benchmarking the main scRNA-seq preprocessing alternatives, including a computational validation and a comprehensive characterization of their divergences, we observed that Kallisto outperforms other methods in the detection and quantification of lncRNAs, due to the latter having less accurate annotation. Expanding on this exhaustive benchmarking, we have developed a specialized workflow, termed ELATUS, to streamline the identification of functionally relevant lncRNAs previously undetected in 10x scRNA-seq experiments. Importantly, experimental validations identified the lncRNA exclusively-detected by ELATUS, *AL121895.1*, as a *cis*-repressor specific of triple negative breast cancer cells. These results underscore ELATUS's potential in uncovering expression patterns of cell type-specific and biologically relevant lncRNAs typically overlooked by standard scRNA-seq pipelines. Finally, the developed workflow, ELATUS is openly available as an R package to facilitate its adoption by the broader biomedical community.

Results

Preprocessing choices strongly affect lncRNA detection by scRNA-seq

A significant number of the RNAs expressed in mammalian cells are transcribed from lncRNA genes[15,49]. The annotation of lncRNAs is constantly evolving[6,15], rendering their quantification more challenging compared to protein-coding genes (Supplementary Fig. 1). A tailored scRNA-seq workflow could shed light on their contribution to individual cell identity, which is largely understudied due to the technical limitations of single-cell technologies in terms of quantification depth and sparsity. We set to evaluate different steps of the scRNA-seq computational pipeline to identity the most suitable analysis for the detection of functional lncRNAs. Due to the more unprecise annotation of lncRNAs, we hypothesized that the quantification model choice could have a strong effect on their detection at single-cell level, as suggested by previous work of our group in bulk RNA-seq[50].

We first conducted a comprehensive benchmarking of current state-of-the-art scRNA-seq preprocessing pipelines, including both the alignment-based methods Cell Ranger and STARsolo, and the pseudoalignment-based methods Kallisto-Bustools and Salmon-Alevin, to evaluate how they affect the detection and quantification of lncRNAs (Fig. 1a). To this end, we first used widely characterized 10x Genomics datasets: 1k brain cells from an E18 mouse[51], and 10k healthy human peripheral blood mononuclear cells (PBMCs)[52]. These public datasets have been already applied for comparing the distinct pipelines[29,34,45,47,48]. In agreement with previous research[47,48], Kallisto produced a slightly higher mapping rate (Fig. 1b and Supplementary Fig. 2a). Besides, the pseudoalignments-based methods presented the shortest running times (Supplementary Fig. 2b and Supplementary Fig. 3a) and were less memory-expensive (Supplementary Fig. 2c and Supplementary Fig. 3b).

We next conducted a quality-control step on each raw cell-by-gene UMI count matrix generated by all pipelines and filtered them for empty droplets. In human PBMCs, the mitochondrial content was very similar in all tested pipelines (Supplementary Fig. 3c), while it was higher with Salmon in the mouse brain dataset (Supplementary Fig. 2d). We then removed cells with high mitochondrial composition and potential multiplets to preserve high-quality cells[28]. We observed that these cells had comparable expression levels among pipelines (Fig. 1c and Supplementary Fig. 2e) and that the majority of them were commonly retained by all of them (Fig. 1d and Supplementary Fig. 2f). Furthermore, the main cell types were distinguished by all tested preprocessing options when using canonical markers (Fig. 1e, Supplementary Fig. 2g, h, Supplementary Fig. 3d, Supplementary Data 1).

Interestingly, regarding gene detection, we observed important differences across pipelines. While the distribution of detected protein-coding genes per cell was more similar, the number of identified lncRNAs per cell by Kallisto was strikingly higher (Fig. 1f, g, Supplementary Fig. 2i, j). To exclude the possibility that these differences were caused by poorly expressed genes with practically no counts, we next retained only those that fulfilled minimal expression thresholds (see "Methods"). While most highly-expressed protein-coding genes were commonly detected by different pipelines (Fig. 1h, Supplementary Fig. 2k), a very significant number of highly-expressed lncRNAs were only recognized by Kallisto, both in human and mouse datasets, whereas the remaining pipelines did not quantify them (Fig. 1i, Supplementary Fig. 2i). Of note, the impact of the preprocessing choice remained when controlling for expression levels (Supplementary Fig. 3e) as well as across an increasing set of expression thresholds (Supplementary Fig. 3f, g), indicating that differences in detection are not due to the lower expression of lncRNAs.

To further investigate whether the observed differences were maintained across different models and tissues, we expanded the benchmarking and analyzed a large and diverse set of public 10x Genomics scRNA-seq datasets. Specifically, we used data from human healthy intestine[53], healthy lung and pulmonary fibrotic samples[54,55], as well as PBMCs from mouse[56] and human (5k cells)[57] (Supplementary Fig. 4). Due to the similarities between the results yielded by the gold-standard pipeline Cell Ranger and the other preprocessing alternatives, with the exception of Kallisto, in what follows, we restricted the assessment to these two preprocessing pipelines.

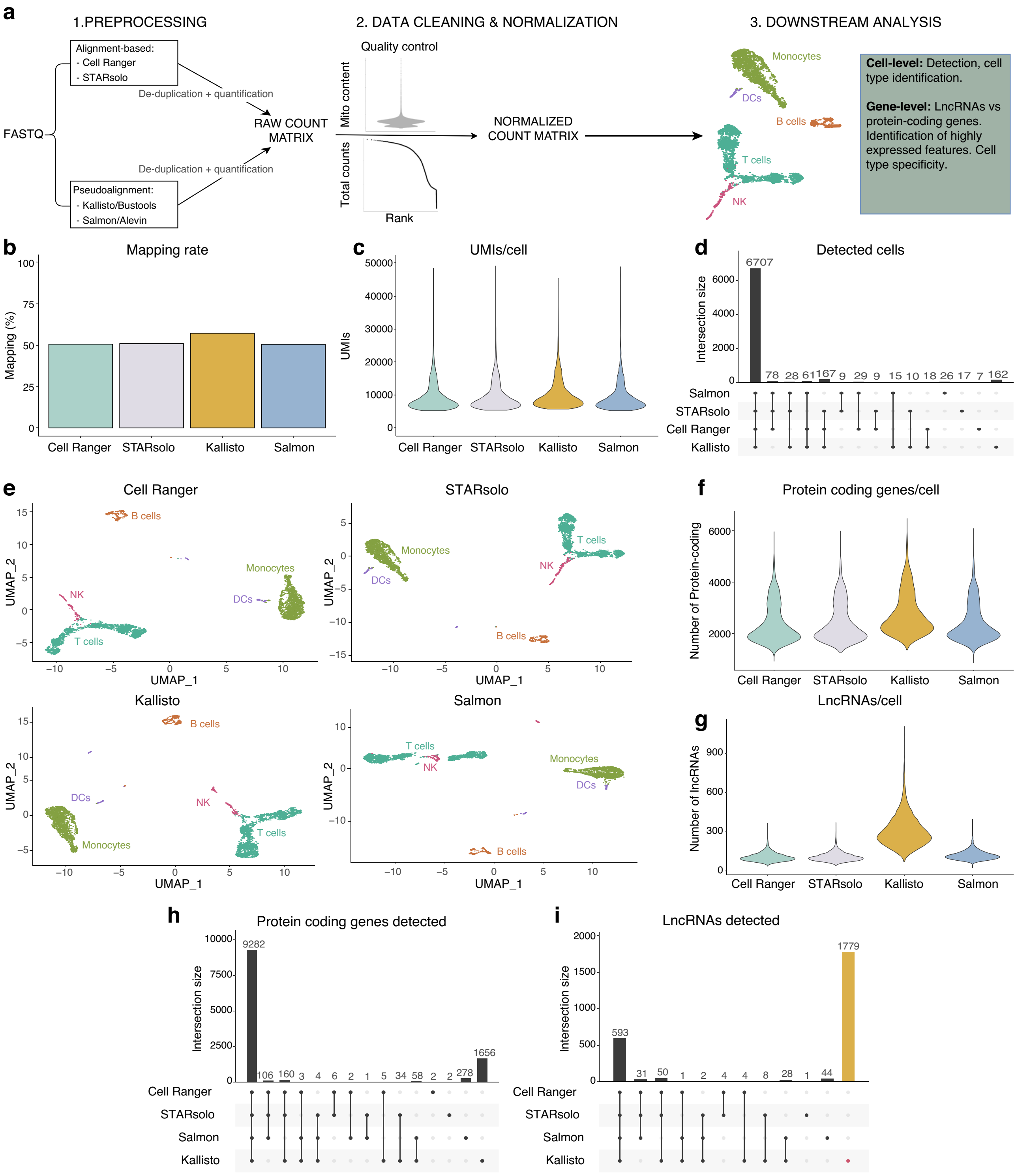

Fig. 1 | Preprocessing choices strongly affect lncRNA detection in a scRNA-seq dataset consisting of 10k human PBMCs from a healthy donor. a Benchmark explanation. Fastq files were preprocessed with the aligner-based Cell Ranger and STARsolo, and the pseudoaligners Kallisto and Salmon. Empty droplets, cells with high mitochondrial content and potential multiplets were filtered, followed by normalization. After dimensionality reduction, clustering and cell type annotation, we compare cell type detection, identification of protein-coding genes and lncRNAs depending on the preprocessing choice. **b** Mapping rate by each pipeline. **c** Number of UMIs per cell by each pipeline. **d** UpSet plot showing the overlap of retained high-quality cells by each pipeline. **e** UMAP plots displaying the main cell types identified across pipelines. **f** Number of detected protein-coding genes per cell across pipelines. **g** Number of detected lncRNAs per cell across pipelines. **h** UpSet plot displaying the overlap of highly-expressed protein-coding genes, (**i**) lncRNAs per pipeline. Only considering genes with more than 250 counts and present in more than 25 cells.

The results on the extended benchmark verified that, after filtering low-quality cells and multiplets, the expression per cell was practically identical across datasets between Cell Ranger and Kallisto (Supplementary Fig. 4a) and most high-quality cells were commonly retained (Supplementary Fig. 4b). With respect to gene detection, while the distribution of protein-coding genes detected per cell was very similar in both pipelines, we confirmed that Kallisto found a remarkably higher number of lncRNAs in each cell (Supplementary Fig. 4c, d). Further, using a gradient of thresholds on expression to filter poorly expressed genes, we corroborated that most protein-coding genes were commonly identified across distinct datasets, whereas there was an important fraction of lncRNAs exclusively captured with Kallisto (Supplementary Fig. 4e, f).

Altogether, these results indicate that, while the detection of mRNAs is not affected, the identification of lncRNAs in scRNA-seq data is severely influenced by the preprocessing choice. In particular, the Kallisto preprocessing pipeline stands out in the detection and quantification of lncRNAs in an expanded and diverse set of scRNA-seq datasets.

scATAC-seq multiome indicates an optimized preprocessing alternative for lncRNA quantification

To assess the biological plausibility of the lncRNAs exclusively quantified by Kallisto, we employed single-cell multiome data. This dataset allows for simultaneous measurement of gene expression via RNA-seq and mapping of open-chromatin using with ATAC-seq within the same cell. We reasoned that scATAC-seq would mirror the expression of lncRNAs identified by scRNA-seq without encountering the same technical biases. Specifically, we selected a public 10x Genomics multiome dataset containing 3k PBMCs from a healthy donor[58] and we tested whether there was more consistency between scATAC-seq profiles and scRNA-seq measurements when the latter was preprocessed with Cell Ranger or Kallisto.

We started by applying general scATAC-seq quality control thresholds (ATAC counts, TSS enrichment and nucleosome signal) to filter low-quality nuclei (Supplementary Fig. 5a–c). Next, we removed nuclei with high mitochondrial content (Supplementary Fig. 5d) and empty droplets (see "Methods"). The expression per nuclei was very similar, although slightly higher with Kallisto (Fig. 2a). Using established markers, we were able to distinguish the main cell types with both scRNA-seq pipelines (Fig. 2b, Supplementary Fig. 5e, Supplementary Data 1). Of note, due to the pre-mRNA reference used in snRNA-seq, Cell Ranger detected more highly-expressed lncRNAs than Kallisto, as well as more lncRNAs per cell (Supplementary Fig. 5f–h). To test the concordance between the scATAC-seq signal and the scRNA-seq gene expression, we constructed a gene activity matrix by counting the scATAC-seq fragments that fall in the gene body and promoter regions. Then, for every high-quality nucleus, we checked which RNA-seq quantification pipeline, Cell Ranger or Kallisto, yielded more genes having coincident ATAC-seq signal and RNA-seq expression (Fig. 2c, see "Methods"). To contemplate diverse scenarios, a gradient of thresholds on both ATAC-seq and RNA-seq was used to classify a gene as simultaneously activated (see "Methods"). Interestingly, for the majority of thresholds, we observed that when scRNA-seq was analyzed by Kallisto, scATAC-seq detected a significant increase in the number of genes simultaneously activated in each nucleus, compared to scATAC-seq coupled to scRNA-seq analyzed by Cell Ranger (Fig. 2d, p-value < 0.0001).

Moreover, the number of nuclei that had more genes coincidently activated according to RNA-Seq and ATAC-Seq was, in general, remarkably higher for Kallisto than Cell Ranger across all thresholds (Fig. 2e, Supplementary Fig. 5f). The improved association between scATAC-seq and scRNA-seq was illustrated in the protein-coding gene *CYP2F1* and the lncRNA *AC242960.3*, where their ATAC-seq profiles across distinct cell types only corresponded to the RNA-seq expression when processed with Kallisto (Fig. 2f).

Together, these results demonstrate that the lncRNAs detected by Kallisto correspond better to the scATAC-seq measurements on the same cells, associated with open and transcriptionally active chromatin, confirming that this is an optimized alternative to improve the quantification of lncRNAs.

Exclusive and commonly identified lncRNAs share similar characteristics

In order to build up a tailored workflow for lncRNA quantification in scRNA-seq, we first needed to delve into the extra-detected lncRNAs to determine whether they could be potential bona fide lncRNAs. To that end, we investigated distinct properties of genes both exclusively and commonly identified by Kallisto (for simplicity termed "exclusive" and "common" genes, respectively), focusing on their expression profiles and sequence composition. In particular, we inspected their absolute expression and length, as well as their repeat content, k-mer distribution, and cell type specificity levels. This was conducted in each scRNA-seq dataset previously included in this work, in which we applied a gradient of thresholds on expression to filter poorly-expressed genes.

We observed that both exclusive lncRNAs and protein-coding genes were significantly less expressed than the common ones (Fig. 3a, Supplementary Fig. 6a). Regarding their length, we noted interesting discrepancies between lncRNAs and protein-coding genes. While exclusive protein-coding genes were significantly longer for every dataset, length differences between exclusive and common lncRNAs were not significantly different for most datasets and under varied thresholds on expression (Fig. 3b, Supplementary Fig. 6b).

Given that the number of exons has been associated with the functionality of lncRNAs[10], we next investigated the distribution of exons that are proximal to the 3′ ends, since the scRNA-seq datasets analyzed in this work are 3′-biased due to the library amplification protocol (Supplementary Fig. 7a). Interestingly, we found that exclusive lncRNAs have, in general, a significantly higher number of exons proximal to the 3′ end than the commonly detected lncRNAs (Fig. 3c, Supplementary Fig. 7b). This contrasts with protein-coding genes, where common protein-coding genes have a significantly higher number of exons near the 3′ end (Fig. 3c, Supplementary Fig. 7b).

Furthermore, it has been noted that pseudoalignments methods could cause poor quality alignments to low-complexity sequences that result in unexpected high expression of particular genes[47]. To test this possibility, we next compared the percentage of each gene that was covered by repeat elements in both exclusively detected and common features and represented this as a ratio. In agreement with this explanation, we found that the repeat content of exclusive protein-coding genes was clearly larger than that of the common ones (Fig. 3d, right). However, the repeat content of exclusive and common lncRNAs was very similar (Fig. 3d, left) suggesting that the increased detection of lncRNAs by Kallisto was not caused by poor quality alignments to low-complexity sequences. In fact, the ratio between the repeat content of exclusive to common protein-coding genes was significantly higher than the ratio between the repeat content of exclusive to common lncRNAs (p-value = 4e-5, one-tailed paired t-test). Similar results were observed when explicitly assessing the influence of transposable elements (Supplementary Fig. 7c).

Aside, it has been documented that lncRNAs with similar k-mer profiles share related functions[59]. We applied this reasoning and analyzed the functional communities according to k-mer profiles. The analysis did not identify communities preferentially composed of exclusively captured lncRNAs. Indeed, functionally-related communities were formed by both common and exclusively identified lncRNAs (Supplementary Fig. 7d), suggesting that exclusive lncRNAs

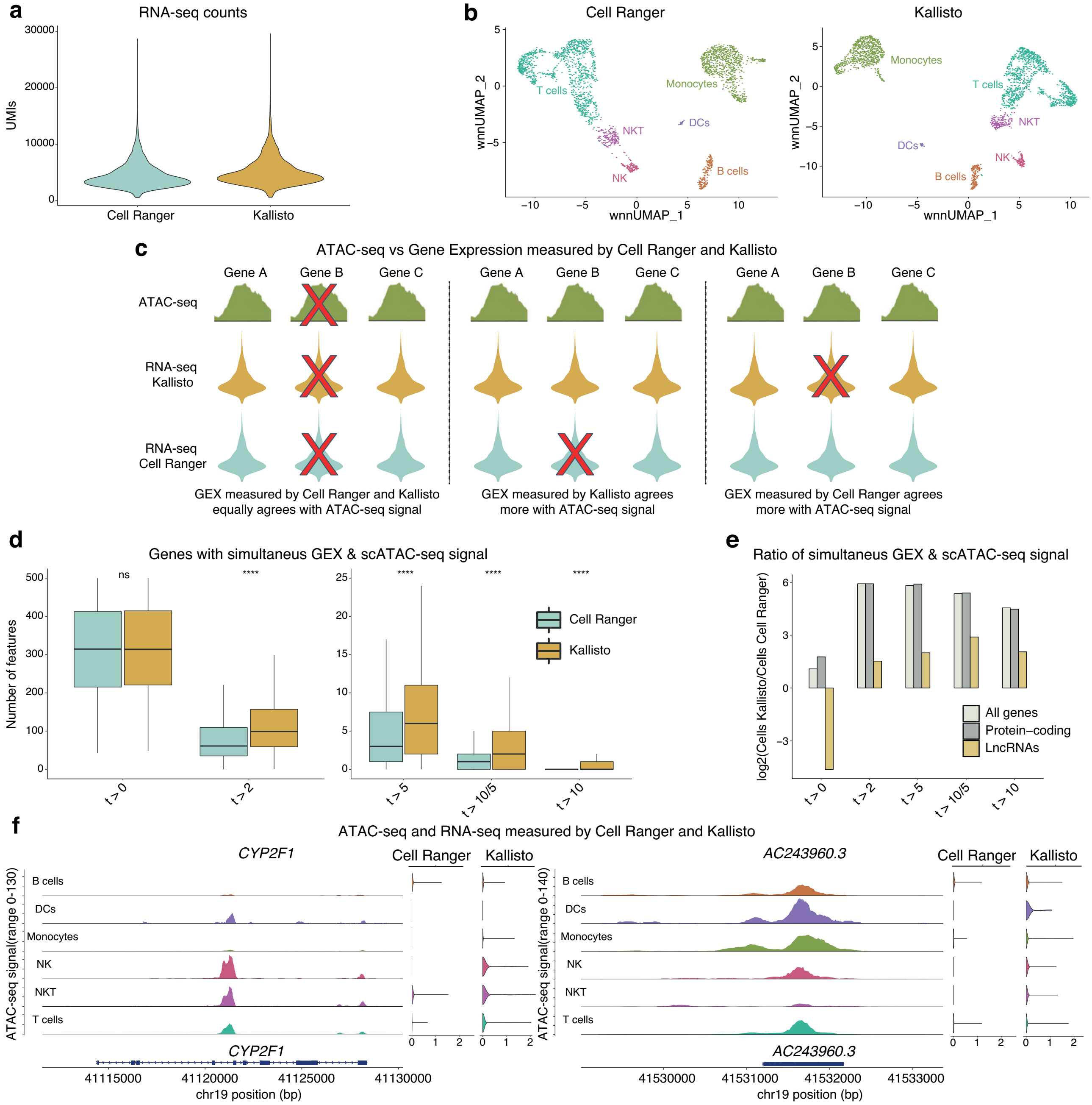

Fig. 2 | scATAC-seq multiome indicates an optimized preprocessing alternative for lncRNA quantification. a Number of UMI counts per cell obtained when preprocessing the scRNA-seq with Cell Ranger and Kallisto. **b** Weighted nearest neighbors UMAP plot displaying the populations of cell types identified by Cell Ranger and Kallisto. **c** Methodology used for comparing the similarity between the scRNA-seq, processed with Cell Ranger or Kallisto, and the Gene Activity matrix obtained from the scATAC-seq data. Specifically, for each nucleus, we count the number of simultaneously expressed genes in the scRNA-seq (with higher expression than a threshold), both with Cell Ranger and Kallisto, and in the Gene Activity matrix (with higher signal than a threshold). **d** Boxplot displaying, for each nucleus ($n = 2538$), the number of simultaneously activated genes when the scRNA-seq is processed with Cell Ranger and Kallisto. Two-tailed student t-test for assessing differential expression. *ns* represents *p*-value > 0.1, * represents *p*-value ≤ 0.1, ** represents *p*-value ≤ 0.05, *** represents *p*-value ≤ 0.005 and **** represents *p*-value ≤ 0.0005. Boxplots represent 25 to 75 percentiles, whiskers are 1.5 x interquantile range (interquantile range = percentile75−percentile25) **e** Ratio of the number of nuclei for which there is more genes simultaneously activated with Kallisto divided by the number of nuclei for which there is more genes simultaneously activated with Cell Ranger. For each nucleus we have considered the expression of all genes (white), only protein-coding genes (gray) and only lncRNAs (yellow). In (**d**) and (**e**), the x-axis represents the different thresholds used for quantifying only a gene as simultaneously activated if it had: (t > 0) at least 1 UMI in RNA-seq and 1 read in ATAC-seq and, (t > 2) at least 3 UMIs in RNA-seq and 3 reads in ATAC-seq, (t > 5) at least 6 UMIs in RNA-seq and 6 reads in ATAC-seq, (t > 10/5) at least 11 UMIs in RNA-seq and 6 reads in ATAC-seq and (t > 10) at least 11 UMIs in RNA-seq and 11 reads in ATAC-seq. **f** ATAC-seq signal and RNA-seq expression, with both Cell Ranger and Kallisto, of protein-coding gene *CYP2F1* (left) and lncRNA *AC243960.3* (right).

are not enriched in a particular function related to k-mer content but rather are comparable to known functional lncRNAs in this regard[59].

Besides, given that lncRNAs are defined to be more cell type-specific than protein-coding genes[7–10], we wondered whether this was maintained for the exclusive lncRNAs. Therefore, we calculated the specificity of the exclusive lncRNAs and compared it with the specificity of protein-coding genes. For this purpose, we defined a specificity index (SI, see “Methods”) that measured how localized or ubiquitous

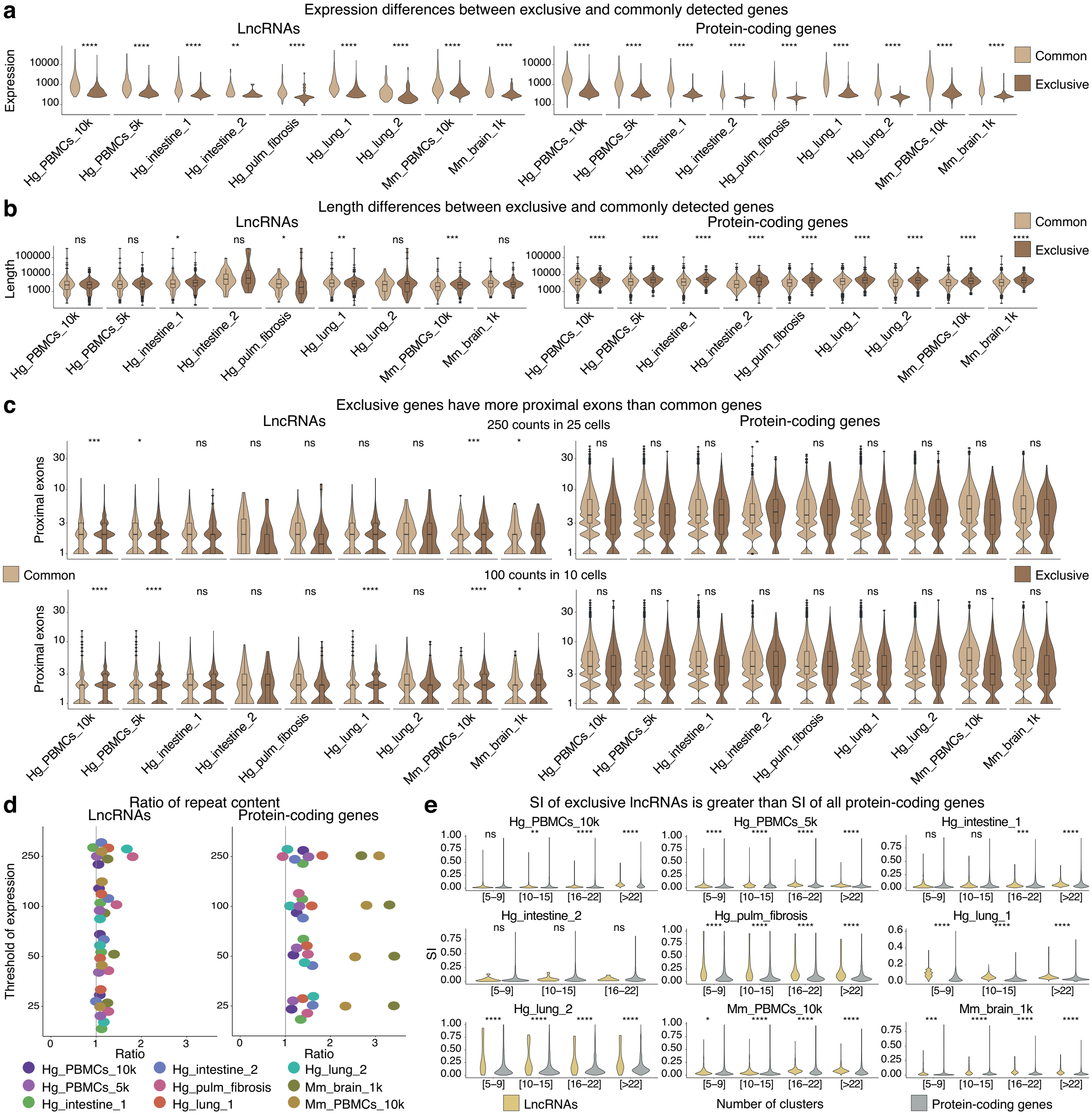

Fig. 3 | Exclusive and commonly identified lncRNAs share similar characteristics. a Normalized expression differences, **b** Length differences and **c** Differences in the number of proximal exons (< 15 kb from the 3′ UTR) of (left) exclusive vs. common lncRNAs (right) exclusive vs. common protein-coding genes. In (**a**) and (**b**) only genes with more than 250 counts and present in more than 25 cells were considered and significance was assessed with a two-tailed Wilcoxon test. In (**c**) only genes with more than (up) 250 (down) 100 counts and present in more than (up) 25 (down) 10 cells were considered and significance was assessed with a one-tailed Wilcoxon test, testing if exclusive genes have more proximal exons. In (**a**), (**b**) and (**c**) common and exclusive lncRNAs and common and exclusive protein-coding genes have the following 'n' for each datasets; Hg_PBMCs_10k: 591, 1774, 9273 and 1653; Hg_PBMCs_5k: 424, 1112, 8479 and 1577; Hg_intestine_1: 261, 401, 8058 and 1368; Hg_intestine_2: 23, 11, 1744 and 376; Hg_pulm_fibrosis: 94, 38, 5543 and 712; Hg_lung_1: 404, 966, 9705 and 911; Hg_lung_2: 80, 50, 5365 and 1004; Mm_PBMCs_10k: 256, 372, 8937 and 844 and Mm_brain_1k: 93, 78, 6293 and 564. **d** Ratio of the percentage of the sequence covered by repeats of exclusive to common (left) lncRNAs (right) protein-coding genes. A jitter on the y-axis was included for ease of visualization. Thresholds for removing lowly-expressed genes; more than i) 25, ii) 50, iii) 100 and iv) 250 counts and present in more than i) 3, ii) 5, iii) 10 and iv) 25 cells. **e** SI differences to test if the SI of exclusive lncRNAs is significantly higher (one-tailed Wilcoxon test), than the SI of protein-coding genes. SI distributions were calculated across distinct sizes of clusters (5-9, 10-15, 16-22 and > 22 clusters. Only genes with more than 250 counts and present in more than 25 cells were considered. In (**a**), (**b**), (**c**) and (**e**) *ns* represents *p*-value > 0.1, * represents *p*-value ≤ 0.1, ** represents *p*-value ≤ 0.05, *** represents *p*-value ≤ 0.005 and **** represents *p*-value ≤ 0.0005 and boxplots represent 25 to 75 percentiles, whiskers are 1.5x interquantile range (interquantile range = percentile75 − percentile 25).

the expression of a gene is. To assess the influence of both the size and the number of clusters on the SI, we clustered the scRNA-seq datasets in a gradient from large to small subpopulations. In accordance with their defined properties, we corroborated that the SI exclusive lncRNAs was significantly higher than that of protein-coding genes, both in large and small subclusters in the majority of datasets (Fig. 3e, Supplementary Fig. 7e). In summary, our findings indicate that exclusive lncRNAs are less expressed than common lncRNAs, have more exons near the 3' end and are more cell type-specific when compared to protein-coding genes.

Inaccurate annotation of lncRNAs causes detection differences

Having observed that exclusive and common lncRNAs share comparable biological features, we then delved into the reasons causing the extra-detection of lncRNAs by Kallisto. We first hypothesized that, out of the thousands lncRNAs exclusively detected, a fraction could be false positives caused by the spurious mapping of intronic reads, in line with observations by prior studies[31,47]. To test this hypothesis, we used the above scRNA-seq dataset consisting on 10k human healthy PBMCs[52], and assessed whether the k-mer overlap between intronic regions and exclusively detected genes is higher than the overlap with commonly detected ones. This analysis, would enable us to determine which set of genes was more likely to contain falsely assigned intronic reads. Surprisingly, the overlap with intronic regions was lower for exclusive genes than in the commonly detected ones (Fig. 4a). Specifically, the overlap between intronic regions and exclusive lncRNAs was in average 10% lower compared with the overlap of commonly detected lncRNAs, which suggests that spurious mapping of intronic reads is not a main driver of erroneously detected lncRNAs among exclusive lncRNAs.

We then hypothesized that the reason underlying the increased detection of lncRNAs in Kallisto, is due to Kallisto model being more robust against non-accurate annotations[15,17]. To test this, we investigated whether the precision of lncRNAs references affects the detection differences. Therefore, using the above scRNA-seq dataset consisting on 10k human healthy PBMCs[52], we tested Cell Ranger and Kallisto with the following references: [1]) The less accurate GENCODE hg19 (v19)[60], 2) the originally applied GENCODE hg38 (v37)[61], 3) the more recent and precise GENCODE hg38 (v45), and 4) the NONCODE annotation (v5)[49], which is the most inclusive and integrated lncRNAs collection. In agreement with our hypothesis, the proportion of exclusively detected lncRNAs compared to the proportion of commonly detected lncRNAs was substantially reduced when testing more precise annotations (Fig. 4b). Indeed, the ratio of exclusively-to-common lncRNAs was practically reduced by half with the most inclusive annotation schemes, compared to the less precise reference GENCODE hg19 (Fig. 4b). In contrast, the detection of protein-coding genes, with a more precise annotation, was minimally affected by the reference releases (Fig. 4c). These results indicate that annotation inaccuracies of lncRNAs are likely responsible for their exclusive detection by Kallisto.

To further investigate whether Kallisto but not Cell Ranger, is able to detect truly expressed lncRNAs whose annotation is not totally accurate, we performed a simulation of lncRNAs expression. To this end, we first generated a simulated *ground truth* annotation by adding artificial modifications to the annotation of lncRNAs from the GENCODE "official" annotation. More specifically, we included an additional exon (of median exonic length) in the middle of the largest intron of each transcript, and expanded all exonic boundaries 100 bp in each direction (see "Methods"). Based on this *ground truth* annotation, we simulated 2000 exonic reads belonging to 1500 randomly subsampled lncRNAs, and randomly divided them in 100 cells. We then preprocessed the generated reads with Cell Ranger and Kallisto. Both models were executed using the unmodified reference to determine whether they could detect true expression originated from regions not annotated as exonic in the official unmodified annotation. When comparing the count matrices generated by Cell Ranger and Kallisto against the *ground truth* count matrix, strikingly, the quantification error of Cell Ranger was practically double than that of Kallisto in each of the 10 simulations performed (Fig. 4d). Furthermore, from the 1500 randomly subsampled lncRNAs, Kallisto detected a significatively higher proportion of highly-expressed lncRNAs (Fig. 4d). These results indicate that part of the simulated reads, could only be rescued by Kallisto and to the contrary, Cell Ranger missed the expression originated from regions not accurately defined as exonic.

Furthermore, as we indicated above, the fact that exclusive lncRNAs have more exons near the 3' end, where most expression is, and hence more splicing-junctions that make them more prone to annotation inaccuracies, is coherent with the explanation that exclusive lncRNAs, or at least an important fraction of them, are not well annotated and only Kallisto was able to capture them. These results showed that Kallisto is able to capture truly expressed, but not perfectly annotated lncRNAs that Cell Ranger missed, and indicate the reasons underlying its additional identification of lncRNAs.

Biologically relevant lncRNAs are uncovered by ELATUS

The additional detection of lncRNAs by Kallisto, enabled hypothesis testing over thousands of them. However, limited by the capabilities to experimentally study lncRNAs, there is a need for specialized workflows to select lncRNAs with potential biological significance.

To provide a curated list of lncRNAs likely of having biological importance, we implemented a computational workflow for Elucidating biologically relevant lncRNAs annotated transcripts using scRNA-seq, termed ELATUS. We reasoned that a biologically significant exclusive lncRNA should be expressed over a threshold and should have a highly cell type-specific expression pattern. Indeed, we observed that the majority of exclusive lncRNAs were tissue-specific, reaffirming the well-known specificity of lncRNAs[7] (Fig. 5a, Supplementary Fig. 8a). However, the biological relevance of these lncRNAs remains to be determined. Interestingly, exclusive lncRNAs were enriched among lncRNAs identified by CRISPRi screenings in multiple human cell lines[10] (FDR < 0.05, hypergeometric test) (Supplementary Data 2, Supplementary Fig. 9), indicating their role in supporting cellular functions. Thus, ELATUS, besides retaining robustly expressed lncRNAs detected by both Cell Ranger and Kallisto, was designed to retain exclusive lncRNAs that were highly specific according to restrictive selection thresholds (Fig. 5b, see "Methods"). Notably, ELATUS uncovered 87 cell type-specific and highly-expressed lncRNAs, which added to the 173 lncRNAs that were also undetected by Cell Ranger and were hits in the CRISPRi screenings[10] and to the 2080 highly-expressed commonly detected lncRNAs (Supplementary Data 3), defined a complete collection of 2340 lncRNAs that exhibit characteristics of functional lncRNAs in the diverse set of scRNA-seq datasets analyzed.

Then, to further explore the potential of the ELATUS workflow in identifying lncRNAs with significant biological roles, we decided to investigate their expression in cells from triple-negative breast cancer (TNBCs) tumors. These highly aggressive breast tumors comprise various cell types, although our understanding of their transcriptional identity is still incomplete. We obtained five patient-derived TNBC fresh tumor biopsies and performed scRNA-seq that was then processed by ELATUS using both Cell Ranger and Kallisto (see "Methods"). After removing low-quality cells and poorly-expressed genes, we observed an important fraction of highly-expressed exclusive lncRNAs (1037 in total) in every sample, whereas most protein-coding genes were commonly detected (Fig. 5c), confirming our previous observations.

Next, we integrated the five TNBC samples, and cells were classified into major and minor cell types using canonical markers and reference datasets[62] (Fig. 5d, e, Supplementary Fig. 8b, c,

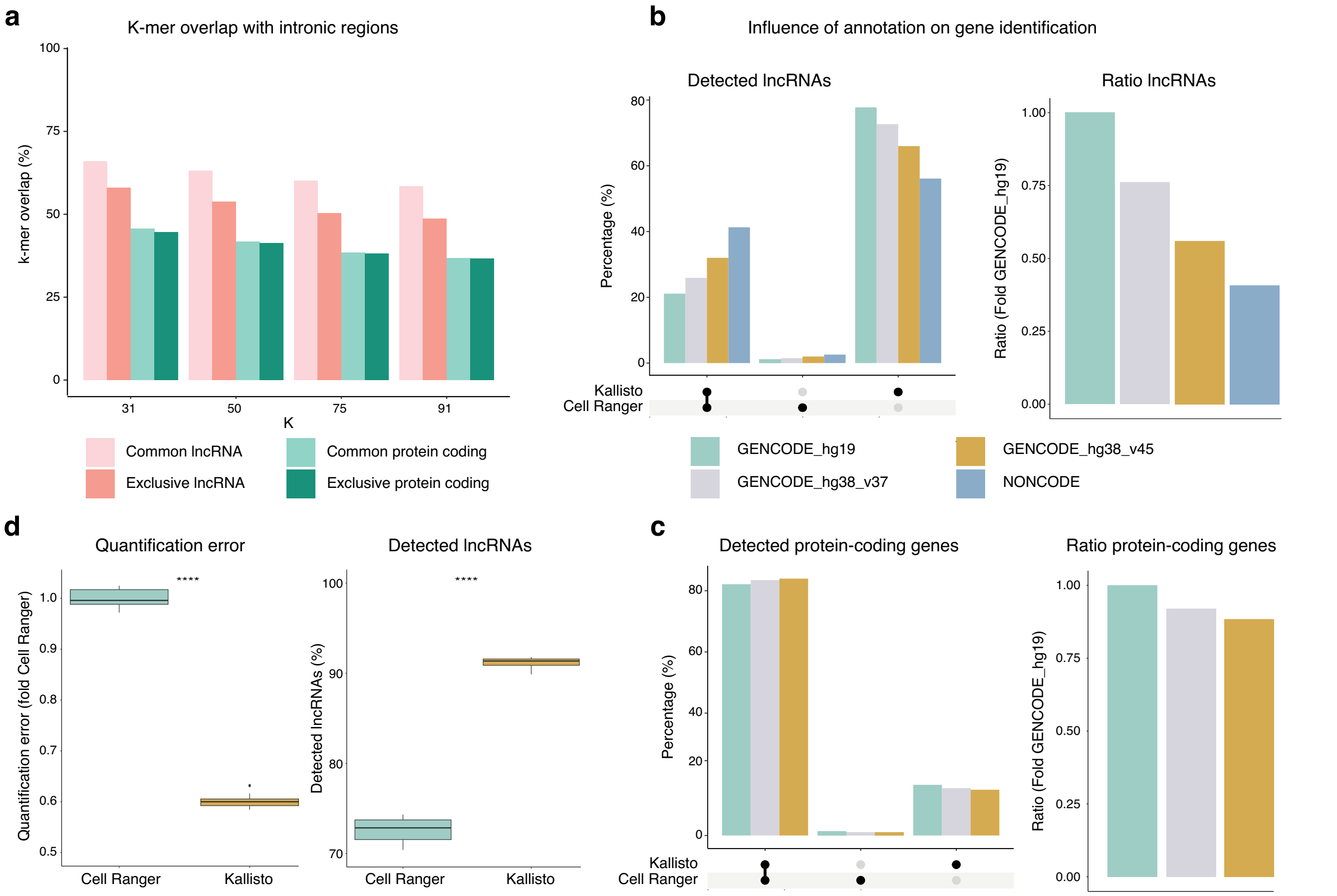

Fig. 4 | Inaccurate annotation of lncRNAs causes detection differences. **a** Percentage of k-mers that overlapped with k-mers of intronic regions with different k-mers lengths. K-mers were generated from the transcript sequences of common and exclusive genes. **b** (left) UpSet plot displaying, for highly-expressed lncRNAs the percentage of them that are detected by Kallisto and Cell Ranger when testing different annotation schemes and (right) ratio of the number of highly-expressed lncRNAs that are exclusively detected by Kallisto divided by the number of highly-expressed lncRNAs that are commonly detected by Cell Ranger and Kallisto. Fold GENCODE hg19. **c** (left) UpSet plot displaying, for highly-expressed protein-coding genes the percentage of them that are detected by Kallisto and Cell Ranger when testing different annotation schemes and (right) ratio of the number of highly-expressed protein-coding genes that are exclusively detected by Kallisto divided by the number of highly-expressed protein-coding genes that are commonly detected by Cell Ranger and Kallisto when testing different annotation schemes. Fold GENCODE hg19. Highly-expressed genes defined as those with more than 250 counts and present in more than 25 cells. Results in (**a**), (**b**) and (**c**) where generated using the scRNA-seq dataset consisting of 10k human PBMCs from a healthy donor. **d** (left) Quantification error between the ground truth matrix with the simulated lncRNA expression and the lncRNAs preprocessing count matrices of Cell Ranger and Kallisto in $n = 10$ simulations. Quantification performed using the Frobenious norm to measure distance between matrices. Quantification errors are normalized to Cell Ranger quantification error (right) Percentage of highly-expressed lncRNAs detected by Cell Ranger and Kallisto in each of the $n = 10$ simulations from the 1500 lncRNAs whose expression is simulated. Highly-expressed genes defined as those with more than 500 counts. Boxplots represent 25 to 75 percentiles, whiskers are 1.5 x interquantile range (interquantile range = percentile75−percentile 25). Statistical significance was assessed with a two-tailed student t-test, *ns* represents *p*-value > 0.1, * represents *p*-value ≤ 0.1, ** represents *p*-value ≤ 0.05, *** represents *p*-value ≤ 0.005 and **** represents *p*-value ≤ 0.0005.

Supplementary Data 1). Further, we observed some lncRNAs overlooked by Cell Ranger that had unique expression patterns (Fig. 5f). Among them, using ELATUS we identified several cell type-specific candidates, such as *WT1-AS* and *AL133679.1*, which are plasmablasts-specific lncRNAs, or *AC009312.1* that is enriched in dendritic cells (Fig. 5g). Interestingly, we also recognized *AL121895.1*, that was specific of breast cancer epithelial cells and could be a potential marker of these tumorigenic cells. These findings highlight the importance of ELATUS for uncovering cell type and cancer type-specific lncRNAs in scRNA-seq experiments.

ELATUS-identified *AL121895.1* is a *cis*-repressor required for triple negative breast cancer progression

Once established that our proposed workflow is able to detect cell type-specific lncRNAs that were previously missed by de facto preprocessing options, we next aimed to experimentally validate their expression patterns in different cell types. To do that, we selected *AL121895.1* and *WT1-AS*, which according to ELATUS are specific to breast cancer epithelial cells and plasmablasts, respectively (Fig. 6a, Supplementary Fig. 8d). We independently analyzed their expression in MDA-MB-231, a human epithelial breast cancer cell line, and in KMS-12-BM, a plasma cell line of multiple myeloma that represents later stages of B-cell differentiation[63–65], similar to plasmablasts. Experimental detection by RT-qPCR confirmed the expression patterns of these lncRNAs that were found by ELATUS, since *AL121895.1* was significantly enriched in breast cancer cells, while *WT1-AS* had a significantly higher expression in multiple myeloma cells (Fig. 6b).

Given that *AL121895.1* is specific to breast cancer epithelial cells, we next investigated its function in these tumoral cells by knocking it down using siRNAs and antisense oligonucleotides (ASOs) (Fig. 6c, Supplementary Fig. 8d). Subcellular fractionation indicated that *AL121895.1* is predominantly present in the chromatin of the cells (Supplementary Fig. 8e). Interestingly, we observed that the knockdown of *AL121895.1* significantly

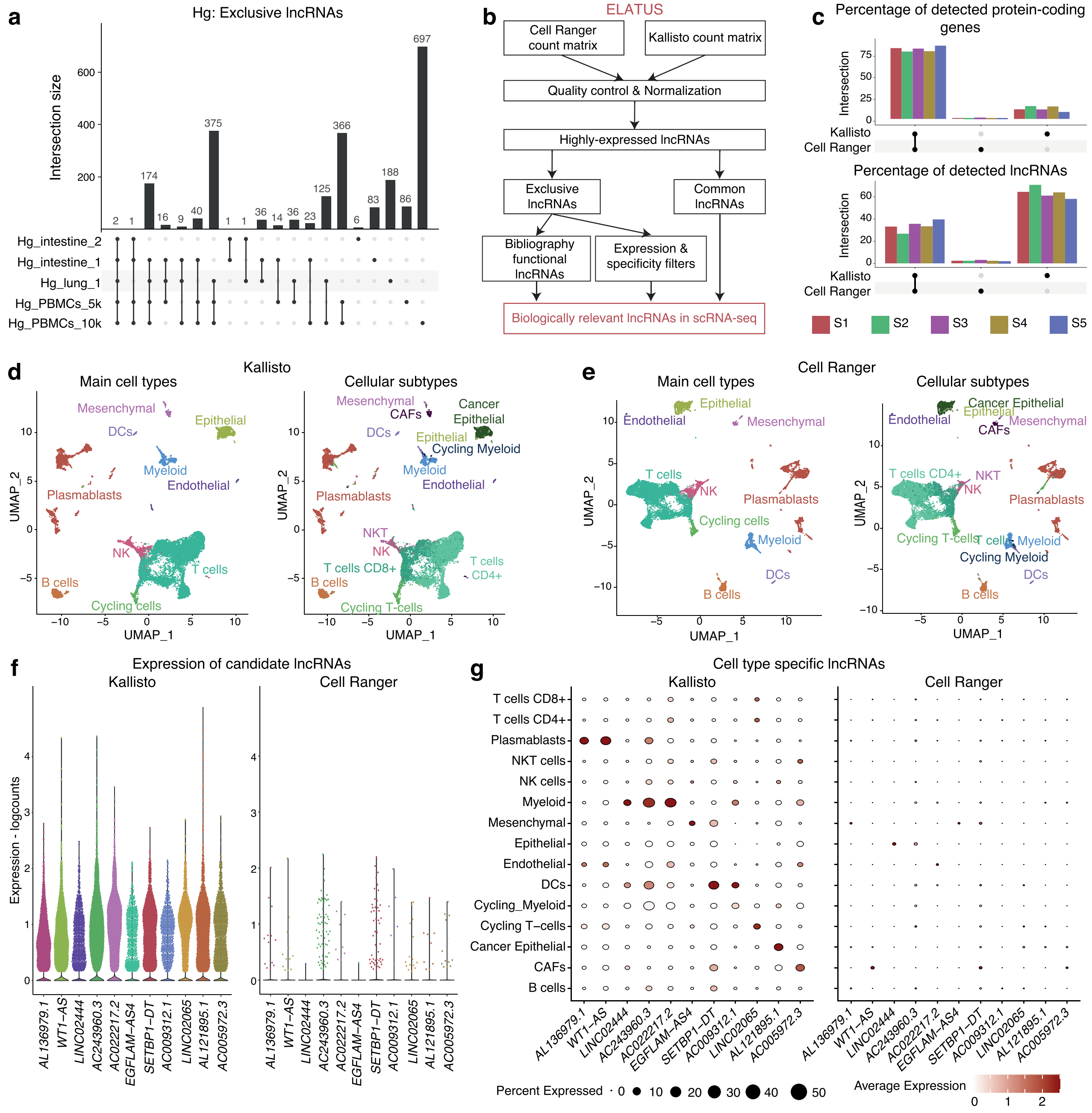

Fig. 5 | Biologically relevant lncRNAs are uncovered by ELATUS. a UpSet plot displaying the overlap of lncRNAs exclusively found by Kallisto in the human scRNA-seq datasets analyzed. **b** ELATUS workflow to uncover biologically important lncRNAs. ELATUS starts importing the raw count matrices obtained with both Cell Ranger and Kallisto. Next, there is a quality control step to distinguish empty droplets from cells, filtering potential multiplets and cells with high mitochondrial content, followed by a normalization and clustering steps. Then, highly-expressed lncRNAs, both commonly detected by Cell Ranger and Kallisto and exclusively detected by Kallisto were selected. All the commonly detected lncRNAs were retained and from the exclusive lncRNAs, ELATUS retained those lncRNAs for which Cell Ranger assigned less than 10 counts, that were 40 times more expressed according to Kallisto than to Cell Ranger and that, according to Kallisto, had a SI > 0.15. ELATUS also retained the exclusive lncRNAs whose functionality has been independently validated by external studies. **c** UpSet plot displaying, as a percentage, the overlap of: left) protein-coding genes, and right) lncRNAs detected by Kallisto and Cell Ranger in each sample. Only genes with more than 250 counts in more than 25 cells were considered in both panels. **d** UMAP plots displaying the different TNBC cell population of: left) main cell types, and right) cell subtypes identified when preprocessing with Kallisto. **e** UMAP plots displaying the different TNBC cell population of: left) main cell types, and right) cell subtypes identified when preprocessing with Cell Ranger. **f** Violin plot showing the expression of some lncRNAs when preprocessing with Kallisto and Cell Ranger. **g** DotPlots showing, with Kallisto and Cell Ranger, the averaged normalized expression in each cellular subtype of these lncRNAs.

decreased the proliferation rate of breast cancer cells (Fig. 6d), indicating a pro-tumorigenic role in these cells. Furthermore, we noticed that its knockdown using siRNAs or antisense oligonucleotides (ASOs) increased the expression of the antisense protein-coding gene, *EPB41L1* (Fig. 6e, Supplementary Fig. 8f), suggesting a previously undocumented *cis*-repressor function for *AL121895.1*. The de-repression of *EPB41L1* was more pronounce when *AL121895.1* was knocked down with the use of ASOs, which are known to deplete RNAs as it is transcribed, supporting a strong component of co-transcriptional regulation. However, the fact that siRNA-mediated knockdown also led to *EPB41L1* de-repression, suggested the implication of an RNA-dependent

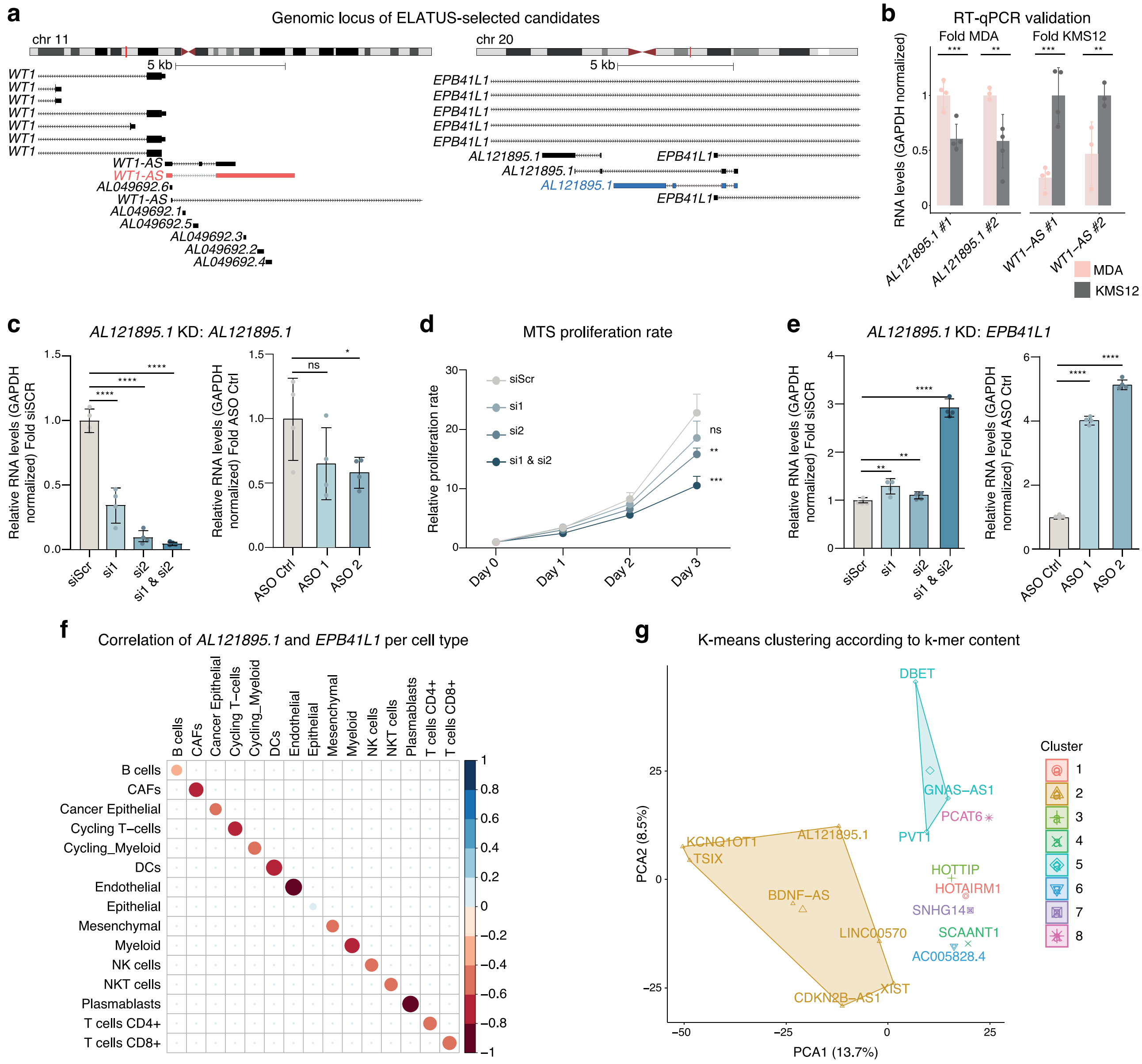

Fig. 6 | ELATUS-identified *AL121895.1* is a *cis*-repressor that participates in triple negative breast cancer progression. a Genomic locus of left) *WT1-AS* and right) *AL121895.1*. In blue and red are represented the isoform of *WT1-AS* and *AL121895.1*, respectively, that contain most scRNA-seq reads assigned by Kallisto. **b** RT-qPCR normalized RNA levels (mean + SD) of *AL121895.1* and *WT1-AS* in MDA and KMS12 cell lines. *AL121895.1* and *WT1-AS* expression has been normalized with respect to MDA and KMS12, respectively. *N* = 4 technical replicates. **c** RT-qPCR normalized RNA levels (mean + SD) showing the expression of *AL121895.1* on MDA cells after treating them with (left) scramble (siSCR) or knocked down with the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2) (right) ASO control, or knocking them out with ASO 1 and ASO 2. *N* = 4 technical replicates **d** MTS proliferation assay (mean + SD) of MDA cells measured during three days treating them with scramble (siSCR) or knocked down with the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2). *N* = 3 technical replicates **e** RT-qPCR normalized RNA levels (mean + SD) showing the expression of *EPB41L1* when treating them with (left) scramble (siSCR) or knocked down with the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2) (right) ASO control, or knocking them out with ASO 1 and ASO 2. *N* = 4 technical replicates. In (**b**), (**c**), (**d**) and (**e**) statistical significance was assessed with a two-tailed student t-test, *ns* represents *p*-value > 0.1, * represents *p*-value ≤ 0.1, ** represents *p*-value ≤ 0.05, *** represents *p*-value ≤ 0.005 and **** represents *p*-value ≤ 0.0005. **f** Correlation plot of the normalized expression of *AL121895.1* and *EPB41L1* in each cellular subtype of the TNBC samples preprocessed with Kallisto. **g** Functional classification by SEEKR using K-means clustering to find communities according to k-mer content of *AL121895.1* together with described lncRNAs *cis*-activators and lncRNAs *cis*-repressors.

mechanism rather than pure transcriotional interference. We speculate that *AL121895.1* may help recruit or release specific chromatin factors, favoring the transcription of *EPB41L1*, which as consequence, leads to the increased proliferative capacity of the cancer cells.

Inspection of the expression of *AL121895.1* and *EPB41L1* in the scRNA-seq of the TNBC samples, also confirmed a strong anticorrelation between them in each cell type (Fig. 6f). Further, since lncRNAs with similar k-mer content are functionally related[59], we clustered, by K-means, the k-mer content of *AL121895.1* together with the k-mer composition of already described lncRNAs that are either *cis*-activators or *cis*-repressors[59]. The clustering grouped *AL121895.1* together with six other lncRNAs, from which five were proved *cis*-repressors (Fig. 6g), supporting its repressive role on its neighboring gene. Thus, all the data suggest that *AL121895.1* is a chromatin-associated lncRNA that acts as a repressor of its antisense gene *EPB41L1* in TNBC cells.

Together, these results demonstrate the biological significance of lncRNAs missed by Cell Ranger and highlight the need for ELATUS, the presented optimized scRNA-seq workflow in order to unlock cellular features encoded by lncRNAs.

Discussion

It has been widely established that lncRNAs regulate gene expression and stability by different mechanisms at multiple levels[5,6,11]. However, given their high cell type specificity[7,8,10], the current knowledge of lncRNAs could be greatly expanded with the application of single-cell technologies. Indeed, prior studies have showed that lncRNAs alone can identify cell types[1,43] and that there are certain cell groups that can only be distinguished by them[3]. Nevertheless, compared to protein-coding genes, the annotation of lncRNAs is less well established and more prone to inaccuracies[15,17], which complicates its correct quantification. In fact, as previously probed in bulk RNA-seq experiments[50], pseudoaligners are promising alternatives to the closed-source and aligner-based Cell Ranger, which is the standard scRNA-seq preprocessing pipeline. In this work we have shown, in a wide and diverse set of public scRNA-seq datasets, that the detection and quantification of lncRNAs is severely affected by the preprocessing choice due their less accurate annotation and that the presented workflow tailored to lncRNAs, ELATUS, is essential, not only for defining an exhaustive collection of functional lncRNAs, but also for uncovering biologically important lncRNAs previously undetected.

Although two previous studies noticed that Kallisto is able to detect a higher number of lncRNAs[29,48], they did not delve into the origin of this behavior. They either noted that it provided a comparable increment in the detection of both lncRNAs and protein-coding genes[48], or attributed it to poorly expressed lncRNAs solely captured with Kallisto, that did not provide a significant biological signal gain[29]. We hypothesize that this increased detection is not happening with Salmon due to its selective alignment strategy, in which the authors recommend including a set of decoy sequences to provide a more precise mapping[37]. Further, preceding investigations have alerted of the reduction in quantification accuracy of pseudoaligners due to the spurious mapping of intronic reads[31,47], which can lead to false expression of particular genes[48]. The developers of Kallisto noted this phenomenon to be plausible but rare[34]. We inspected and determined the reasons that caused this increased detection of lncRNAs and provided multiple evidence supporting the validity of the proposed workflow for the identification of lncRNAs.

Due to the notably greater number of lncRNAs exclusively detected by Kallisto, we aimed to confirm whether they exhibit characteristics consistent with bona fide lncRNAs and to investigate the reasons for this improved detection. To achieve this, we conducted a thorough characterization of the exclusive lncRNAs in comparison to those commonly detected. Importantly, in this study we tested and corroborated that the less accurate annotation of lncRNAs is the main cause of the additional detection of lncRNAs. By comparing different reference schemes and simulating expression from regions not correctly annotated as exonic, we were able to confirm that Kallisto is able to detect truly expressed lncRNAs that are not perfectly annotated. Besides, the fact that exclusive lncRNAs have a significantly higher number of proximal exons supports this hypothesis.

Furthermore, we showed that the overlap with intronic regions is higher for commonly captured lncRNAs, which suggests that the explanation attributing spurious alignments of intronic reads does not fully apply to lncRNAs. However, in order to mitigate the risk of including false transcripts and enable robust inference of lncRNAs, here we implement a computational pipeline, termed ELATUS, which enriches for lncRNAs with functional features. ELATUS generates a collection of biologically relevant lncRNAs by retaining, not only those robustly detected lncRNAs, but also, by filtering those that exhibit characteristics of functional lncRNAs from the thousands exclusively identified by Kallisto.

To determine the validity of ELATUS, the different expression patterns were tested computationally and experimentally. Indeed, the expression of lncRNAs computed by Kallisto correlated more closely with ATAC-seq data than that of those detected by Cell Ranger. This indicates that pseudolignment effectively identified bona fide transcripts generated from regions of open chromatin. Moreover, the proposed computational workflow, ELATUS (available online as an R package), in addition to preserving robust and commonly detected lncRNAs, it unveils highly cell-type specific and biologically relevant lncRNAs from among the thousands of exclusive candidates missed by Cell Ranger.

It is of particular relevance the previously uncharacterized lncRNA *AL121895.1*, specific of breast cancer epithelial cells and whose functionality at single-cell level could not be determined by standard preprocessing. Indeed, *AL121895.1* acts as a *cis*-repressor lncRNAs regulating *EPB41L1* expression and promoting TNBC progression. *EPB41L1* encodes a multifunctional protein that has been shown to mediate interactions between the erythrocyte cytoskeleton and the overlying plasma membrane, although it is also expressed in other tissues[66,67]. Moreover, *AL121895.1* is associated with TP53 mutations, a main hallmark of cancer[68], according to the PDAClncDB database[69].Our results indicate that the regulation of *EPB41L1* by *AL121895.1* occurs at the chromatin level, and is dependent on the RNA product of *AL121895.1*, not just its transcription. These findings could have clinical implications, since *AL121895.1* could potentially be an actionable cancer vulnerability, targetable by the newly evolving antisense drugs. The presented data evidences the potential impact of ELATUS to unveil important biological roles of lncRNAs and to expand the map of interactions, in individual cell populations, between the expression of a previously undetected by Cell Ranger lncRNA and the nearby protein-coding gene. Moreover, it exemplifies how scRNA-seq can inform mechanistic questions, such as the *cis* vs *trans* regulatory roles of lncRNAs.

It should be noted that the library preparation method could also influence the detection of lncRNAs. Conventional 10x Genomics scRNA-seq library preparation protocols target only polyA transcripts. Since an important fraction of lncRNAs are not polyadenylated[6], these technologies cannot achieve a complete map of the transcriptome. Therefore, alternative scRNA-seq preparation protocols that capture both polyadenylated and non-polyadenylated transcripts, processed following the ELATUS workflow could potentially reveal an expanded number of functional non-coding transcripts that participate in important cellular functions.

Another direction that should be explored is the improvement of the reference annotation, since annotation inaccuracies explain a significant part of ELATUS better performance. On the one hand, the application of an intronic reference, already recommended for snRNA-seq and that considers both mature and unmature RNAs has been widely discussed in the scRNA-seq community. However, the pre-mRNA reference poses a disjunctive when an exon of a gene is overlapped by an intron of another gene since every read falling in that region would be considered ambiguous (Supplementary Fig. 10a, b) without reaching a clear consensus on how to resolve these situations[70–72]. On the other hand, tissue-specific de novo annotation can greatly increase the detection of poorly annotated transcripts structures of specific lncRNAs[43]. However, generating such annotations requires a high use of resources, and still may not recapitulate highly cell type-specific transcript forms, for which ELATUS can still remain a useful alternative. Future investigation combining an improved reference annotation together with ELATUS could provide significant improvements in gene detection, especially for less studied biotypes such as lncRNAs.

Finally, with the proposed workflow, we favor the detection of lncRNAs with higher cell type specificity, as defined by the high specificity index (SI). We reason that this set of lncRNAs will include those with the most interesting biological features. However, while high cell specificity is recognized as a general characteristic of lncRNAs[7–10], the existence of ubiquitous lncRNAs playing essential roles[73,74] should not be excluded. Here, we propose an optimized computational workflow for analyzing scRNA-seq experiments that has the potential to unlock cellular features and transcriptional complexity, increasing insights into cell identity and lncRNA biology.

Methods

Single-cell RNA-seq preprocessing pipelines

The following scRNA-seq preprocessing pipelines were benchmarked: Cell Ranger, STARsolo, Kallisto-Bustools (referred as Kallisto) and Salmon-Alevin (referred as Salmon). All pipelines were executed with the default recommended parameters in the user guides. For scRNA-seq analysis, Cell Ranger count was run in version 3.0.1, STAR in version 2.7.9, Kallisto in version 0.46.1, Bustools in version 0.40.0 and Salmon in version 1.4.0. For the pulmonary fibrosis dataset, since it was necessary to split cDNA sequence in more than one file, Kallisto was run in version 0.46.2 following authors indications[75].

Reference annotation and generation of the indexes

Human and mouse reference genome, transcriptome and annotation were downloaded from GENCODE[13]. Particularly, for human, hg38 (v37), and for mouse, mm39 (v27), were selected. We created the indexes for each preprocessing pipeline following recommended settings. Specifically, Cell Ranger and STARsolo were indexed against the entire genome, Kallisto against the transcriptome and Salmon was indexed to perform selective alignment to the transcriptome with full decoys as suggested by both the authors and independent benchmark studies[31,47,48,76]. The commands for preprocessing and generating the indexes for each preprocessing tool can be found in https://github.com/ML4BM-Lab/manuscript_scRNAseq_lncRNAs.

To analyze the evolution in the number of annotated protein-coding genes and lncRNAs, we analyzed the annotation provided by GENCODE since version 7, published in 2010, until the version 44, published in December 2022.

scRNA-seq public datasets

Different scRNA-seq prepared with 10x Genomics protocols, with both v2 and v3 chemistries, were used. Most of them were public, while sequencing data regarding TNBC samples were manually prepared. The description of each dataset as well as their link to access the sequencing data is provided in Supplementary Data 4.

scRNA-seq quality control, gene detection and post-processing steps

Raw count matrices were used to standardize preprocessing pipelines as input for quality control, where we followed common scRNA-seq computational guidelines[77]. Specifically, emptydrops[78] was applied to distinguish empty droplets from cells in each dataset processed with Cell Ranger, STARsolo and Kallisto. On the other side, Salmon, employs a whitelisting filtering strategy to filter empty droplets and it does not output the raw count data. To account for that and standardize filtering strategies, the minimum number of counts surviving emptydrops filtering in Cell Ranger, STARsolo and Kallisto was selected as an additional threshold to filter cells in Salmon with fewer counts than this defined threshold. Potential doublets were then identified and removed with scDblFinder[79]. Finally, cells with high mitochondrial content and an abnormally high number of counts were also filtered.

Once low-quality cells had been removed, the detection and quantification of protein-coding genes and lncRNAs were compared. In order to test the differences between lncRNAs and protein-coding genes by the overall lower expression of lncRNAs, we calculated the total expression of each lncRNA with each preprocessing pipeline (for simplicity termed "lncRNAs overall expression"). Then, we only considered those genes less expressed than the mean "lncRNAs overall expression" plus i) 0.5 ii) iii) 2 * the standard deviation of the lncRNAs overall expression.

Next, poorly expressed genes were also filtered by applying a gradient of thresholds on the expression. The different thresholds applied retained those genes with more than 1) 250 counts and present in more than 25 cells, 2) 100 counts and present in more than 10 cells, 3) 50 counts and present in more than 5 cells and 4) 25 counts and present in more than 3 cells. These thresholds were also applied in the characterization of the genes exclusively identified by Kallisto compared to genes commonly found ones. Due to the differences in the preserved number of cells in each scRNA-seq dataset, both absolute numbers and percentages have been used to represent the differences and the overlap of the detected highly-expressed genes by Kallisto and Cell Ranger.

After the quality control was completed, normalization was performed using logNormCounts function from scuttle R package[80]. Dimensionality reduction was conducted using runPCA, runTSNE and runUMAP functions from scater R package[81]. Next, clustering was performed on a generated shared nearest-neighbor (SNN)[82] graph using the Louvain community detection algorithm to cluster the cells[83,84]. These clusters were manually annotated to cell types using canonical markers (Supplementary Data 1). Subtypes in TNBC integrated data were distinguished using the annotation program JIND[85] selecting as a reference a dataset consisting of ~45000 cells from TNBC samples[62]. We assessed the correlation (Spearman) between the normalized expression of *AL121895.1* and *EPB41L1* per these cellular subtypes in each cell that expressed either *AL121905.1* or *EPB41L1* (or both).

Single-cell multiome analysis

In 3k PBMCs sequenced with single cell multiome, scATAC-seq raw data has been directly downloaded from 10X Genomics website[58] on which scATAC-seq specific thresholds were first applied to remove low-quality nuclei. In particular, nuclei with very few or excessive ATAC-seq counts were filtered, as well as those with high nucleosome signal or little enrichment at the TSS[86].

Gene expression data was obtained with Cell Ranger count (version 5.0.1) using—include-introns option. Regarding Kallisto (version 0.46.1) the index was generated to include both introns and exons using—workflow lamanno parameter.

For the raw RNA-seq matrices, those nuclei the fit the scATAC-seq quality control thresholds in both Cell Ranger and Kallisto were retained. Nuclei with very few RNA-seq counts or very high mitochondrial content were further removed. Next, ATAC-seq data from these high-quality nuclei were normalized using a Latent Semantic Indexing approach. "Weighted nearest neighbor" (WNN) analysis was then performed to integrate the ATAC-seq with the gene expression obtained by Cell Ranger and Kallisto[87]. This was performed by normalizing the ATAC-seq applying term frequency inverse document frequency (TF-IDF) and Singular Value Decomposition (SVD) and utilizing FindMultiModalNeighbors function from Seurat using default parameters (20 mutimodal neighbors to compute) following developers' recommendations[88]. This integrated data was used for generating the clusters (following indications in scRNA-seq post-processing section) that were manually assigned to different cell types according to the expression of established marker genes.

To compare the similarity between the ATAC-seq signal and the RNA-seq gene expression, the GeneActivity function from Signac R package[89] was applied to obtain a gene activity matrix by counting the scATAC-seq fragments that fall in each gene body (+ 2Kb upstream from the TSS). Then, for every high-quality nucleus we compared (student's t-test) the number of genes that have simultaneous ATAC-seq-signal and RNA-seq expression when preprocessed with Cell Ranger or with Kallisto. A gene was defined to be simultaneously activated if its ATAC-seq signal and its RNA-seq expression were higher than a gradient of defined thresholds. Specifically, from less to more restrictive thresholds; if they had at least 1) 1 read in ATAC-seq and 1 UMI in RNA-seq, 2) 3 reads in ATAC-seq and 3 UMIs in RNA-seq, 3) 6 reads in ATAC-seq and 6 UMIs in RNA-seq, 4) 6 reads in ATAC-seq and 11 UMIs in RNA-seq and 5) 11 reads in ATAC-seq and 11 UMIs in RNA-seq.

The differences, per cell, were also represented as an odds ratio (in log2 scale) showing the likelihood of having more genes simultaneously activated with Kallisto than with Cell Ranger. Further, the ratio (in log2 scale), of the number of nuclei for which there were more simultaneous activation when the snRNA-seq was processed with Kallisto than with Cell Ranger has been computed. A positive ratio indicates a better correspondence between ATAC-seq and Kallisto than between ATAC-seq and Cell Ranger, while a negative ratio indicates the opposite.

Exclusive vs. common genes: Length, proximal exons, repeat & TE content and k-mer analysis

To investigate the length, the number of exons that are closer than 15 kb from the 3' UTR and repeat content of both exclusive and commonly detected genes in every scRNA-seq dataset, the longest isoform of each gene was selected. The annotation of repeats for both human and mouse genomes was downloaded from RepeatMasker (version 4.1.5)[90], where we considered all distinct types of repeats with the exception of microsatellites repeats. We calculated the ratio between the percentage of the sequence of exclusive features covered by repeats divided by the percentage of the sequence of common features covered by repeats. The annotation of transposable elements was downloaded from the Hammel lab[91] and we calculated the ratio between the percentage of sequence of exclusive features covered by transposable elements divided by the percentage of sequence of common features covered by transposable elements.

We compared the k-mer content of all lncRNAs using SEEKR[59], a software developed for sequence evaluation through k-mer content, where we calculated the 6-mer functionally-related communities using the canonical isoform of each gene following the default recommended settings by the authors[92]. For the K-means (stats R package) clustering of *AL121895.1* according its k-mer content and the k-mer content of described cis-activators (DBET, HOTAIRM1, HOTTIP, LINC00570, PCAT6, PVT1) or cis-repressors (BDNF-AS, CDKN2B-AS1, KCNQ1OT1, TSIX, XIST, GNAS-AS1, SNHG14, SCAANT1)[59], we considered the isoform *AL121895.1-EST00000441208*. This is the isoform to which Kallisto assigned most of scRNA-seq reads and for which RT-qPCR primers and siRNAs were designed.

Exclusive vs. common genes: Specificity index

In order to implement the Specificity Index (SI), in line with other methods[4,43], each scRNA-seq datasets was clustered across distinct ranges of number and sizes of clusters, from fewer and bigger clusters to more but smaller clusters. We started by splitting the cells in 5-9 clusters, then 10–15 clusters, then 16-22 clusters and finally generating a very detailed subdivision with at least 23 clusters. The SI metric was then designed in order to inform if a lncRNA is more specific of big or small subpopulations. We implemented the SI following the Shannon-Entropy specificity (HS) formulation defined in TSPEX, a library with several specificity metrics[93]. So, in order to define the SI for each gene *g*, we first calculated its mean expression x_i in each cluster, $i = 1,2...n$, where n is the number of clusters. Next, we calculated for each gene, the proportion of mean expression in each cluster, P_i:

$$P_i = \frac{x_i}{\sum_{i=1}^{n}(x_i)} \tag{1}$$

Finally, using the entropy HS formulation, we implemented the SI metric, where we assessed if each gene is expressed in fewer and localized clusters or if its expression is more broadly expressed. Concretely the SI, for each gene, was defined as:

$$SI = 1 + \sum_{i=1}^{n} P_i{}^{*}\log_{n} P_i \tag{2}$$

The SI is ranked from [0–1], from genes very ubiquitously expressed to very cluster-specific genes. A gene whose expression is equally distributed across different clusters will have a SI of 0, while if a gene is exclusively expressed in one cluster its SI will be 1.

Bulk RNA-seq analysis from human healthy PBMCs

Bulk RNA-seq public data from healthy human PBMCs were analyzed from sample GSM3172785[94]. Raw fastq files were preprocessed with Kallisto and STAR in order to generate the count matrices and compare gene expression in bulk and scRNA-seq.

Intronic regions: K-mer analysis

To determine whether Kallisto exclusive or commonly detected genes could be originated from spurious mapping of intronic regions, we assessed the k-mer overlap between intronic regions and the transcript sequences of commonly and exclusively detected genes. Starting with k = 31, the default k-mer size applied by Kallisto, we tested larger k-mer sizes. Specifically, we tested k = 50, k = 75 and k = 91, which was the dataset's read length.

Simulation of lncRNAs expression in scRNA-seq

ScRNA-seq reads have been extracted from exonic regions indicated in the official reference annotation and from artificial exonic modifications in order to include probable unannotated exonic regions[17,95,96], considering that lncRNAs are universally alternatively spliced[97]. Every exon was expanded 100 bp in each direction and an exon in the middle of the largest intron of each transcript was added (with the median exonic length). This constituted a simulated synthetic *ground truth* annotation that could represented lncRNAs whose annotation is not correctly reflected by the GENCODE "official" reference.

Based on this *ground truth* annotation, we simulated expression from every exonic region of 1500 randomly subsampled lncRNAs. We assigned 2000 reads to each of them and randomly divided in 100 cells. This process generated a *ground truth* expression matrix. Then, we created the corresponding fastq files from the simulated hastq files from the simulated reads, that were preprocessed with Cell Ranger and Kallisto using the "official" unmodified reference. After performing 10 simulations, the quantification error between the preprocessed count matrixes generated by Cell Ranger and Kallisto and the simulated ground truth matrixes was compared using the Frobenious norm[98].

ELATUS workflow defines a collection of functional lncRNAs

On the one side, to get the list of biologically relevant lncRNAs, we analyzed the 288 CRISPR functionally validated lncRNAs in multiple human cell lines[10]. Using a hypergeometric test with FDR correction[99], we tested the significance of their overlap with those lncRNAs exclusively identified by Kallisto, in every human scRNA-seq. In total, there was an overlap of 173 Kallisto-exclusive lncRNAs.

On the other side, we implemented ELATUS in order to capture the highly-expressed lncRNAs commonly detected and to inspect the highly-expressed lncRNAs exclusively identified with Kallisto to assess their biological relevance. Therefore, we started by importing the raw count matrices obtained with both Cell Ranger and Kallisto and we performed emptydropts to distinguish empty droplets from cells. Next, we removed potential multiplets and performed a quality-control filtering, followed by a normalization and clustering steps. Further, we integrated samples from the same tissue. Then, highly-expressed lncRNAs (i.e. those with more than 250 counts and present in more than 25 cells), both commonly detected by Cell Ranger and Kallisto and exclusively identified by Kallisto were selected. All the commonly detected lncRNAs were retained, whereas from the exclusive lncRNAs ELATUS retained those lncRNAs for which Cell Ranger assigned lass than 10 counts, that were 40 times more expressed according to Kallisto than to Cell Ranger and that, according to Kallisto, had an SI > 0.15. ELATUS also retained the exclusive lncRNAs whose functionality has been independently validated by external studies. To include a representative set of cluster sizes to calculate the SI, scRNA-seq datasets were divided in different cluster sizes, from 10 to 19 clusters (Supplementary Data 4). ELATUS, which is openly available as an R package in https://github.com/ML4BM-Lab/ELATUS, has been executed with these restrictive thresholds to ensure biological relevance and to minimize the risk of false expression caused by spurious mapping.

Statistical analysis and data plotting

Post-processing analysis were performed in R (version 4.1.2). Barplots and violin plots were represented with ggplot2 (v.3.4.2), where ggpubr (v.0.4.0) was used to test statistical test significance. The specific statistical test for each analysis is detailed in its figure caption, where *ns* represents *p*-value > 0.1, * represents *p*-value ≤ 0.1, ** represents *p*-value ≤ 0.05, *** represents *p*-value ≤ 0.005 and **** represents *p*-value ≤ 0.0005.

From Seurat (v.4.0.1)[87], DimPlot function was used to plot UMAP dimensionality reduction plots and DotPlot and FeaturePlots functions were applied to evaluate gene expression in different cell types and dimensionality reduction spaces, respectively. UpSet plots were generated with UpSetR (v.1.4.0)[100] and ggupset (v.0.3.0). In the analysis of single-cell multiome data, Signac R package (v. 1.9.0)[89] was applied to create the coverage and expression plots.

Cell lines and growth conditions

MDA-MB-231 (HTB-26) were purchased from the American Type Culture Collection (ATCC). MDA-MB-231 cells were cultured in DMEM (GIBCO), supplemented with 10% fetal bovine serum (GIBCO) and 1x penicillin/streptomycin (Lonza). KMS-12-BM cells were from Xabier Agirre's lab at CIMA, University of Navarra. KMS-12-BM cells were grown in RPMI-1640 (GIBCO) medium with 20% fetal bovine serum (GIBCO), 1% penicillin, and 2% Hepes. All of them were maintained at 37˚C and 5% CO_2.

RNAi

For RNA knockdown, siRNAs, which were designed using the i-Score designer tool and purchased from Sigma. Antisense oligonucleotides, with the same targeting sequence, were synthesized by iDT, with 3'-o-methoxyethyl nucleotides on the 5' and 3' ends, as well as consecutive oligodeoxynucleotides to support RNaseH activity (Supplementary Table 1). Control ASO was designed and synthesized bi Ionis Pharmaceuticals.

MDA-MB-231 cells were transfected with Lipofectamine 2000 (Invitrogen) in Serum-free Opti-MEM (GIBCO), following manufacturer instructions. siRNAs and ASOs were transfected for 24 hours at 40 nM and 50 nM final concentrations respectively.

Proliferation assay

Cell proliferation was measured using the CellTiter96 Aqueous Non-Radioactive Cell Proliferation Assay (MTS) kit (Promega). After 24 hours transfection, 1000 MDA cells were cultured in M-96 plate wells. Every 24 hours, 20 μL of MTS reagent (Promega) were added to culture media, and incubated for 2 h, prior to 490 λ measurement. Triplicate measures were normalized to day 0, and statistical differences between control and experimental conditions at day 3 were calculated with a two-tailed student t-test.

Cellular fractionation

For cellular fractionation, all steps were performed in the presence of protease inhibitors (Roche), phosphatase inhibitors (Roche), and RNAsin (Promega). A total of 1 × 107 MDA-MB-231 cells were harvested with trypsin, washed, and resuspended in 200 μl of isotonic lysis buffer (10 mM Tris-HCl pH7, 150 mM NaCl, 0.15% NP-40) for 5 min on ice and layered on a sucrose buffer (10 mM Tris-HCl, 150 mM NaCl, 25% sucrose). Nuclei were centrifuged for 10 min at 13.000 × *g*, to recover the supernatant as the cytoplasmic fraction. The nuclear pellet was washed (1 mM EDTA, 0.1% Triton-X100 in PBS), and resuspended in 200 μl glycerol buffer (20 mM Tris-HCl pH8, 75 mM NaCl, 0.5 mM EDTA, 505 glycerol, 0.85 mM DTT) and finally lysed with 200 μl of nuclear lysis buffer (20 mM HEPES, 300 mM NaCl, 1 M urea, 0,2 mM EDTA, 1% NP-40, 1 mM DTT). Lysed nuclei were centrifuged at 13.000 × *g* for 2 min to separate the soluble fraction (supernatant) from the chromatin-associated fraction (pellet).

RNA extraction, processing, and RT-qPCR

Cell preparations were fixed with TRIzol (Sigma), and RNA precipitated with isopropanol. RNA extraction was followed by Turbo DNAse (Invitrogen) digestion for 30 minutes at 37 °C. For RT-qPCR, 1 μg RNA was reverse-transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied 30 Biosystem) with random hexamer primers, following manufacturer instructions. The obtained cDNA was analyzed by quantitative PCR (qPCR) using iTaq Universal SYBR Green supermix (Bio-Rad) in a ViiA™ 7 Real-Time PCR System machine (ThermoFisher), all reactions performed in quadruplicate. For total extract analysis, GAPDH RNA levels were used for normalization. To assess subcellular RNA distribution, relative RNA levels found in chromatin, nuclear and cytoplasmic extracts were represented as a percentage of a whole, where relative *GAPDH* and *MALAT1* levels were used as control cytoplasmic and chromatin-localized RNAs, respectively. *U6* and *U4* snRNAs were used as nuclear control RNAs. Statistical differences between relative RNA levels were calculated by unpaired two-tailed Student's *t*-test. RT-qPCR primers were self-designed or designed with the NCBI Primer designing tool, and purchased from Metabion (Supplementary Table 2).

Human samples

All patients participating in the study provided written informed consent. The study and the use of all clinical materials have been approved by the Research Ethics Committee of the University of Navarra under decision number 2021.058Mod1. Samples and data from patients included in the study were provided by the Biobank of the University of Navarra and were processed following standard operating procedures approved by the Ethical and Scientific Committees.

TNBC sample preparation and preparation of the scRNA-seq libraries

Five biopsies of tissue from different patients of breast cancer were processed following manufacturer's instructions of Human Tumor Dissociation Kit from Miltenyi Biotec. Briefly, 2-3 tissue cylinders per patient were cut into small slices of 2–4 mm, digested with enzymes H, R and A, and dissociated with gentleMACS Dissociator (Miltelnyi Biotec). After 30 minutes incubation at 37 °C and a short centrifugation step (400 × *g*, 1 minute at 4 °C), sample material at the bottom of the tube was collected. cell suspension was then moved to a Falcon cell strainer (70 μm) placed on a 50 mL tube, washed with 20 mL of DMEM, and centrifuged at 400 × *g* for 5 minutes at 4 °C. Pelleted cells were washed with 1 ml of PBS 1×0.05%BSA, supplemented with 5ul RNAse OUT (Invitrogen), and transferred to a Dolphin tube (Sorenson) where cell suspension was centrifuged at 400 × *g* for 5 minutes at room temperature. Viability (> 70%) of resuspended cells was corroborated with cellometer (Nexcelom).

The transcriptomes of 16,000-20,000 cells were examined using Single Cell 3' Reagent Kits v3.1 (10X Genomics) according to the manufacturer's instructions. Briefly, 17000–20000 cells were loaded at a concentration of 1000 cells/μL on a Chromium Controller instrument (10X Genomics) to capture single cells in gel bead-in-emulsions (GEMs). In this step, each cell was encapsulated with primers containing a fixed Illumina Read 1 sequence, a cell-identifying 16 bp 10x Genomics barcode, a 12 bp Unique Molecular Identifier (UMI) and a poly-dT sequence. Upon cell lysis, reverse transcription yielded full-length, barcoded cDNA. This cDNA was then released from the GEMs, PCR-amplified and purified with magnetic beads (SPRIselect, Beckman Colter). Enzymatic Fragmentation and Size Selection was used to optimize cDNA size prior to library construction. Fragmented cDNA was then end-repaired, A-tailed and ligated to Illumina adapters. A final PCR-amplification with barcoded primers allowed sample indexing. Library quality control and quantification was performed using Qubit 3.0 Fluorometer (Life Technologies) and Agilent's 4200 TapeStation System (Agilent), respectively. Sequencing was performed in a NextSeq2000 (Illumina) (Read1: 28; Read2: 91; i7 index: 8) at an average depth of 500000 reads/sample.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the findings of this study are available from the corresponding authors upon request. TNBC scRNA-seq data have been publicly stored in NCBI with the following identifier: GSE246142. Public scRNA-seq data of human intestine was studied from GSM480339 and GSM480348. Public scRNA-seq data of human lung was downloaded from GSM4037320, GSM5020383, GSM4037316. Human public scRNA-seq of PBMCs and mouse public scRNA-seq datasets were downloaded from 10x Genomics website. Human public bulk RNA-seq from PBMCs was downloaded from GSM3172785.

Code availability

The code and scripts to reproduce the results are freely available and can be accessed in https://github.com/ML4BM-Lab/manuscript_scRNAseq_lncRNAs[101]. ELATUS R package for elucidating biologically relevant lncRNA annotated transcripts using scRNA-seq is available at https://github.com/ML4BM-Lab/ELATUS.

References

1. Rahman, R. U. et al. Singletrome: a method to analyze and enhance the transcriptome with long noncoding RNAs for single cell analysis. https://doi.org/10.1101/2022.10.31.514182.
2. Luo, H. et al. Single-cell long non-coding RNA landscape of T cells in human cancer immunity. *Genomics Proteom. Bioinforma.* **19**, 377–393 (2021).
3. Zheng, L. L. et al. ColorCells: a database of expression, classification and functions of lncRNAs in single cells. *Brief. Bioinform* **22**, 1–11 (2021).
4. Santus, L. et al. Single-cell profiling of lncRNA expression during Ebola virus infection in rhesus macaques. *Nat. Commun. 2023 14:1* **14**, 1–14 (2023).
5. Statello, 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).
6. Mattick, J. S. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).
7. Cabili, M. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
8. Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex https://doi.org/10.1186/s13059-016-0932-1 (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. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (1979)* **355**, aah7111 (2017).
11. Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* **21**, 1253–1261 (2015).
12. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
13. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
14. Hezroni, H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* **11**, 1110–1122 (2015).
15. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* **19**, 535–548 (2018).
16. Kornienko, A. E. et al. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.* **17**, 1–23 (2016).
17. Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
18. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 1–4 (2020).
19. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
20. Prescott, S. L., Umans, B. D., Williams, E. K., Brust, R. D. & Liberles, S. D. An airway protection program revealed by sweeping genetic control of vagal afferents. *Cell* **181**, 574–589.e14 (2020).
21. La Manno, G. et al. RNA velocity of single cells. *Nat.* **560**, 494–498 (2018).
22. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
23. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
24. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).
25. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
26. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2017).
27. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **53**, 1005–1005 (2021).
28. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
29. You, Y. et al. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biol.* **22**, 339 (2021).
30. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

31. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.05.442755 (2021).
32. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
33. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
34. Melsted, P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39**, 813–818 (2021).
35. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. *Bioinformatics* **35**, 4472–4473 (2019).
36. Srivastava, A., Malik, L., Smith, T., Sudbery, I. & Patro, R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.* **20**, 1–16 (2019).
37. Srivastava, A. et al. Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.* **21**, 1–29 (2020).
38. See, K. et al. Single cardiomyocyte nuclear transcriptomes reveal a lincRNA-regulated de-differentiation and cell cycle stress-response in vivo. *Nat. Commun.* **8**, 1–13 (2017).
39. Kim, D. H. et al. Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell* **16**, 88–101 (2015).
40. Hu, W., Wang, T., Yang, Y. & Zheng, S. Tumor heterogeneity uncovered by dynamic expression of long noncoding RNA at single-cell resolution. *Cancer Genet.* **208**, 581–586 (2015).
41. Johnsson, P. et al. Transcriptional kinetics and molecular functions of long noncoding RNAs. *Nat. Genet.* **54**, 306–317 (2022).
42. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
43. Bitar, M. et al. Redefining normal breast cell populations using long noncoding RNAs. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkad339 (2023).
44. He, Z. et al. Single-cell transcriptome analysis dissects lncRNA-associated gene networks in Arabidopsis. *Plant Commun.* **5**, 100717 (2024).
45. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* **10**, 1–11 (2019).
46. Du, Y., Huang, Q., Arisdakessian, C. & Garmire, L. X. Evaluation of STAR and Kallisto on single cell RNA-Seq data alignment. *G3 Genes Genomes Genet.* **10**, 1775–1783 (2020).
47. He, D. et al. Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nat. Methods* **19**, 316–322 (2022).
48. Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S. & John, D. Comparative analysis of common alignment tools for single-cell RNA sequencing. *Gigascience* **11**, giac001 (2022).
49. Fang, S. et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **46**, D308–D314 (2018).
50. Zheng, H., Brennan, K., Hernaez, M. & Gevaert, O. Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. **8**, 1–13 (2019).
51. 1k Brain Cells from an E18 Mouse (v3 chemistry) – 10x Genomics. https://www.10xgenomics.com/resources/datasets/1-k-brain-cells-from-an-e-18-mouse-v-3-chemistry-3-standard-3-0-0.
52. PBMCs from a Healthy Donor: Whole Transcriptome Analysis - 10x Genomics. https://www.10xgenomics.com/resources/datasets/pbm-cs-from-a-healthy-donor-whole-transcriptome-analysis-3-1-standard-4-0-0.
53. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution ll Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826 (2021).
54. Schupp, J. C. et al. Integrated single-cell atlas of endothelial cells of the human lung. *Circulation* **144**, 286–302 (2021).
55. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).
56. 10k Mouse PBMCs Multiplexed, 2 CMOs - 10x Genomics. https://www.10xgenomics.com/resources/datasets/10-k-mouse-pbm-cs-multiplexed-2-cm-os-3-1-standard-6-0-0.
57. 5k Peripheral Blood Mononuclear Cells (PBMCs) from a Healthy Donor (Next GEM) - 10x Genomics. https://www.10xgenomics.com/resources/datasets/5-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-next-gem-3-1-standard-3-0-2.
58. PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting (3k) - 10x Genomics. https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-2-0-0.
59. Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet.* **50**, 1474–1482 (2018).
60. GENCODE - Human Release 19. https://www.gencodegenes.org/human/release_19.html.
61. GENCODE - Human Release 45. https://www.gencodegenes.org/human/release_45.html.
62. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
63. Namba, M. et al. Establishment of five human myeloma cell lines. *Vitr. Cell. Developmental Biol.* **25**, 723–729 (1989).
64. Edwards, J. C. W. & Cambridge, G. B-cell targeting in rheumatoid arthritis and other autoimmune diseases. *Nat. Rev. Immunol.* **6**, 394–403 (2006).
65. Jourdan, M. et al. An in vitro model of differentiation of memory B cells into plasmablasts and plasma cells including detailed phenotypic and molecular characterization. *Blood* **114**, 5173–5181 (2009).
66. Wang, H. et al. Selective effects of protein 4.1N deficiency on neuroendocrine and reproductive systems. *Sci. Rep.* **10**, 1–14 (2020).
67. Kim, A. C., Van Huffel, C., Lutchman, M. & Chishti, A. H. Radiation hybrid mapping ofEPB41L1,a novel protein 4.1 homologue, to human chromosome 20q11.2–q12. *Genomics* **49**, 165–166 (1998).
68. Petitjean, A., Achatz, M. I. W., Borresen-Dale, A. L., Hainaut, P. & Olivier, M. TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* **26**, 2157–2165 (2007).
69. AL121895.1. https://www.maherlab.com/pdaclncdb/al121895.1.
70. Hjörleifsson, K. E., Sullivan, D. K., Holley, G., Melsted, P. & Pachter, L. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. https://doi.org/10.1101/2022.12.02.518832.
71. He, D., Soneson, C. & Patro, R. Understanding and evaluating ambiguity in single-cell and single-nucleus RNA-sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.04.522742 (2023).
72. Pool, A. H., Poldsam, H., Chen, S., Thomson, M. & Oka, Y. Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references. *Nat. Methods* **20**, 1506–1515 (2023).
73. Chakraborty, S. et al. Harnessing the tissue and plasma lncRNA-peptidome to discover peptide-based cancer biomarkers. *Sci. Rep.* **9**, 1–17 (2019).
74. Goyal, B. et al. Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer. *BBA-Rev. Cancer* **1875**, 188502 (2021).

75. SC5P-R2 sequencing · Issue #226 · pachterlab/kallisto. https://github.com/pachterlab/kallisto/issues/226.
76. Selective Alignment. https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/.
77. Amezquita, R. A. et al. Orchestrating single-cell analysis with bioconductor. *Nat. Methods* **17**, 137–145 (2019).
78. Lun, A. T. L. et al. EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 1–9 (2019).
79. Germain, P. L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research* **10**, 979 (2021).
80. LTLA/scuttle: Clone of the Bioconductor repository for the scuttle package. https://github.com/LTLA/scuttle/.
81. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
82. Lun, A. T. et al. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, 2122 (2016).
83. Network Analysis and Visualization [R package igraph version 1.5.1]. (2023).
84. igraph – Network analysis software. https://igraph.org/.
85. Goyal, M. et al. JIND: joint integration and discrimination for automated single-cell annotation. *Bioinformatics* **38**, 2488–2495 (2022).
86. Joint RNA and ATAC analysis: 10x multiomic • Signac. https://stuartlab.org/signac/articles/pbmc_multiomic.
87. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
88. Weighted Nearest Neighbor Analysis • Seurat. https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.
89. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
90. RepeatMasker Home Page. https://www.repeatmasker.org/.
91. Index of /shares/mhammelllab/www-data/TEtranscripts/TE_GTF. https://labshare.cshl.edu/shares/mhammelllab/www-data/TEtranscripts/TE_GTF/.
92. CalabreseLab/seekr: A library for counting small kmer frequencies in nucleotide sequences. https://github.com/CalabreseLab/seekr.
93. Camargo, A. P., Vasconcelos, A. A., Fiamenghi, M. B., Pereira, G. A. G. & Carazzolle, M. F. tspex: a tissue-specificity calculator for gene expression data. 1–7 https://doi.org/10.21203/RS.3.RS-51998/V1 (2020).
94. Zucca, S. et al. RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls. *Sci. Data* **6**, 1–8 (2019).
95. Zhang, J. et al. Deep annotation of long noncoding RNAs by assembling RNA-seq and small RNA-seq data. *J. Biol. Chem.* **299**, 105130 (2023).
96. Melé, M. et al. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).
97. Deveson, I. W. et al. Universal alternative splicing of noncoding exons. *Cell Syst.* **6**, 245–255.e5 (2018).
98. Böttcher, A. & Wenzel, D. The Frobenius norm and the commutator. *Linear Algebra Appl.* **429**, 1864–1885 (2008).
99. Benjaminit, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
100. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
101. Goñi, E. et al. Uncovering functional lncRNAs by scRNA-seq with ELATUS. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.26.577344 (2024).

Acknowledgements

We thank Dr. Xabier Agirre for providing the KMS-12-BM samples. We particularly acknowledge the patients for their participation and the Biobank of the University of Navarra for its collaboration. This work was supported by grants to: M.He: RYC2021-033127-I and PID2023-151980OB-I00 financed by MCIN/AEI/10.13039/501100011033. H2020 Marie S. Curie IF Action, European Commission, Grant Agreement No. 898356. M.Hu: La Caixa Foundation [LCF/PR/HR21/00176], European Research Council Consolidator 771425 and Worldwide Cancer Research grants 20-0204. P.F: Blanca (0011-1411-2021-000070). E.G. was supported by PhD fellowship from Gobierno de Navarra (0011-0537-2020-000038).

Author contributions

E.G., designed and performed most analyzes. A.M. performed experimental validations. J.G. performed subcellular fractionations. A.A. generated sequencing data from TNBC patients. M.S. collected clinical samples from TNBC patients. P.F. supervised, E.G., in the scRNA-seq analysis of the TNBC patients. M.Hu. and M.He. conceived the study, supervised the work, and obtained funding. E.G., M.Hu., and M.He. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54005-7.

Correspondence and requests for materials should be addressed to Maite Huarte or Mikel Hernaez.

Peer review information *Nature Communications* thanks Rory Johnson and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://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) 2024