Uncovering functional lncRNAs by scRNA-seq with ELATUS

Authors

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

Affiliations

[1] Center for Applied Medical Research, University of Navarra, PIO XII 55 Ave, 31008, Pamplona, Spain

[2] Institute of Health Research of Navarra (IdiSNA), Pamplona, Spain

[3] Department of Medical Oncology, Clinica Universidad de Navarra, Pio XII 36 Ave, 31008 Pamplona, Spain

[4] Liver and Digestive Diseases Networking Biomedical Research Centre (CIBERehd), Spanish Network for Advanced Therapies (TERAV ISCIII), Madrid, Spain

[5] Cancer Center Clinica Universidad de Navarra (CCUN), Madrid, Spain

[6] Data Science and Artificial Intelligence Institute (DATAI), Universidad de Navarra, Pamplona, Spain

[*] Co-corresponding authors, mhernaez@unav.es & maitehuarte@unav.es

Abstract

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 has been challenging due to their less accurate annotation and low expression compared to protein-coding genes. To identify the important, albeit widely overlooked, highly specific lncRNAs from scRNA-seq data, here we develop a computational framework, ELATUS, based on the pseudoaligner Kallisto that enhances the detection of functional lncRNAs previously undetected and exhibits higher concordance with the ATAC-seq profiles in single-cell multiome data. Importantly, we then independently confirmed the expression patterns of cell type-specific lncRNAs exclusively detected with ELATUS and unveiled biologically important lncRNAs, such as *AL121895.1*, a previously undocumented cis-repressor lncRNA, whose role in proliferation of breast cancer cells was 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.

Introduction

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 the emergence of novel technologies has allowed the profiling of transcriptomes at single-cell resolution, novel computational methods are crucially required.

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 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], 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[16–18] and allowing the inspection of cell dynamics[19]. Particularly, scRNA-seq droplet-based methods[20,21], predominantly the 10x Genomics technology, have revolutionized the procedure by increasing the throughput of cells and decreasing the sequencing costs[22–24].

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[25], which precedes the downstream analysis[26,27]. 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[28], such as the widely used Cell Ranger[22] (developed by 10x Genomics) or STARsolo[29],

perform the entire preprocessing step. In addition, the pseudoaligners Kallisto and Salmon, which are based on matching read k-mers to the transcriptome[30,31], have also incorporated a suite of tools to preprocess the scRNA-seq sequenced reads, named Bustools[32,33] and Alevin[34], respectively.

Despite the differential characteristics of lncRNAs, an exhaustive comparison of scRNA-seq preprocessing pipelines focused on lncRNAs detection and quantification is missing [27,32,34–37]. This comparison is essential to understand their underlying mechanisms and roles at cell resolution.

Here, after benchmarking the main scRNA-seq preprocessing alternatives, including a computational validation and a comprehensive characterization of their divergences, we observed that Kallisto highlights in the detection and quantification of lncRNAs. Expanding on this exhaustive benchmarking, we have developed a specialized workflow, termed ELATUS, to streamline the identification of functionally relevant lncRNAs previously undetected in scRNA-seq experiments. Importantly, experimental validations highlight the potential benefits of ELATUS to discover expression patterns of cell type-specific lncRNAs that are ignored by standard scRNA-seq pipelines and are involved in essential biological processes. Finally, the developed workflow, ELATUS is openly available as an R package to facilitate its adoption by the broad biomedical community.

Results

Preprocessing choices strongly affect lncRNA detection by scRNA-seq

LncRNAs constitute a very significant fraction of the cell transcriptome[38]. The annotation of lncRNAs is constantly evolving[6,15], rendering their quantification more challenging compared to protein-coding genes. 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 hypothesized that the quantification model choice could have a strong effect on the detection of lncRNAs at single-cell level, as suggested by previous work of our group in bulk RNA-seq[39].

We first conducted a comprehensive benchmarking of current state-of-the-art scRNA-seq quantification 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[40], and 10k healthy human peripheral blood mononuclear cells (PBMCs)[41]. These public datasets have been already applied for comparing the distinct pipelines[27,32,35,36,42]. In agreement with previous research[36,42], Kallisto produced a slightly higher mapping rate (**Fig. 1b and Supplementary Fig. 2a**). Besides,

139 in line with previous results[27,32,36,42], the pseudoalignments-based methods
140 presented the shortest running times (**Supplementary Fig. 2b and**
141 **Supplementary Fig. 3a**) and were less memory-expensive (**Supplementary**
142 **Fig. 2c and Supplementary Fig. 3b**), with Kallisto being less computationally
143 intense than Salmon.
144
145 We next conducted a quality-control step on each raw cell-by-gene UMI count
146 matrix generated by all pipelines and filtered them for empty droplets. In human
147 PBMCs, the mitochondrial content was very similar in all tested pipelines
148 (**Supplementary Fig. 3c**), while it was higher with Salmon in the mouse brain
149 dataset (**Supplementary Fig. 2d**). We then removed cells with high
150 mitochondrial composition and potential multiplets to preserve high-quality
151 cells[26]. We observed that these cells had comparable expression levels among
152 pipelines (**Fig. 1c and Supplementary Fig. 2e**) and that the majority of them
153 were commonly retained by all pipelines (**Fig. 1d and Supplementary Fig. 2f**),
154 in consonance with prior studies[27,36]. Furthermore, the main cell types were
155 distinguished by all tested preprocessing options when using canonical markers
156 (**Fig. 1e, Supplementary Fig. 2g-h Supplementary Fig. 3d, Supplementary**
157 **Table 1)**.
158
159 Interestingly, regarding gene detection, we observed important differences
160 across pipelines. While the distribution of detected protein-coding genes per cell
161 was highly similar, the number of identified lncRNAs per cell by Kallisto was
162 strikingly higher (**Fig. 1f-g, Supplementary Fig. 2i-j**). To exclude the possibility
163 that these differences were caused by poorly expressed genes with practically
164 no counts, we next retained only those that fulfilled minimal expression
165 thresholds. While most highly-expressed protein-coding genes were commonly
166 detected by different pipelines (**Fig. 1h, Supplementary Fig. 2k**), a very
167 significant number of highly-expressed lncRNAs were only recognized by
168 Kallisto, both in human and mouse datasets, whereas the remaining pipelines
169 did not quantify them (**Fig. 1i, Supplementary Fig. 2i**). The impact of the
170 preprocessing choice on the detection of lncRNAs was consistent across an
171 increasing set of thresholds on expression (**Supplementary Fig. 3e-f**).
172
173 To further investigate whether the observed differences were maintained across
174 different models and tissues, we expanded the benchmarking and analyzed a
175 large and diverse set of public 10x Genomics scRNA-seq datasets. Specifically,
176 we used data coming from human healthy intestine[43], healthy lung and
177 pulmonary fibrotic samples[44,45], as well as PBMCs from mouse[46] and human (5k
178 cells)[47] (**Supplementary Fig. 4**). Due to the similarities between the results
179 yielded by the gold-standard pipeline Cell Ranger and the other preprocessing
180 alternatives, with the exception of Kallisto, in what follows, we restricted the
181 assessment to these two preprocessing pipelines.
182
183 The results on the extended benchmark verified that, after filtering low-quality
184 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 amount 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 the optimal preprocessing alternative for lncRNA quantification

To investigate whether the lncRNAs exclusively quantified by Kallisto were biologically plausible, we used single-cell multiome data, where in simultaneous cells, gene expression is measured with RNA-seq and open-chromatin is mapped with ATAC-seq. We reasoned that scATAC-seq should orthogonally reflect the expression of lncRNAs identified by scRNA-seq. Specifically, we selected a public 10x Genomics multiome dataset containing 3k PBMCs from a healthy donor[48] and we tested if the scATAC-seq profiles were more consistent with the scRNA-seq measurements when 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. 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 Table 1**). 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**). 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 scATAC-seq coupled to scRNA-seq analyzed by Kallisto give a statistically significant increase in the number of

genes, in each nucleus, that are simultaneously activated, 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 simultaneous cells, associated with open and transcriptionally active chromatin, confirming that this is a better choice to improve the detection of lncRNAs expression.

Exclusive and commonly identified lncRNAs share similar characteristics

We wanted to determine whether Kallisto could be a finer preprocessing method for building up a workflow for lncRNA quantification in scRNA-seq. To that end, we delved into the potential reasons driving its increased detection of lncRNAs. 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, length and number of exons, 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.

In agreement with previous research[27], 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 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. 6b**).

Given that the number of exons is associated with the functionality of lncRNAs[10], we wondered whether its distribution differs between exclusively detected and common genes, as an abnormal exon distribution could indicate annotation errors and therefore explain quantification differences. In general, for all different thresholds applied, the number of exons of exclusive lncRNAs is significantly higher compared to common lncRNAs. This contrasts with protein-coding genes, where there is no evident trend, since in many datasets the

differences are not significant while in others significance appears in any direction (**Fig. 3c, Supplementary Fig. 7a**). The fact that exclusive lncRNAs had more exons, and hence more splicing-junctions that make them more prone to annotation inaccuracies, supported our initial hypothesis that pseudoaligners might be more suited to lncRNA quantification due to the unprecise annotation of lncRNAs[15] compared to that of protein-coding genes.

A particular case has been noted in which pseudoalignment methods output unexpected high expression of the gene *col14a1b* that could be due to poor quality alignments to low-complexity sequences[42]. 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 ratio between the repeat content of exclusive to common protein-coding genes was clearly above one (**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 is significantly higher than the ratio between the repeat content of exclusive to common lncRNAs (p-value 4e-5, one-tailed paired t-test).

Aside, it has been documented that lncRNAs with similar k-mer profiles share related functions[49]. We applied this reasoning and discarded that functional communities according to k-mer profiles were entirely composed of exclusively captured lncRNAs. Indeed, functionally-related communities were formed by both common and exclusively identified lncRNAs (**Supplementary Fig. 7b**), indicating that exclusive lncRNAs are not enriched in a particular function and are likely related to their exon number, as initially hypothesized.

Additionally, 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 the expression of a gene is, in line with other specificity metrics implemented in previous investigations[4,50,51]. To determine the influence of both the size and the number of clusters on the SI, we clustered the scRNA-seq datasets in a gradient from big to small subpopulations. In accordance with their defined properties, we corroborated that the SI exclusive lncRNAs was significantly higher than the SI of protein-coding genes, both in big and small subclusters in the majority of datasets (**Fig. 3e, Supplementary Fig. 7c**).

Our findings indicate that exclusive lncRNAs are less expressed than common lncRNAs, but also highly cell type-specific when compared to protein-coding genes. In terms of sequence composition, both groups of lncRNAs are

comparable, although lncRNAs solely found by Kallisto have more exons. The observations obtained from the performed characterization reinforce its advantages for lncRNA analysis in scRNA-seq, enabling hypothesis testing over vastly more lncRNAs than with other pipelines. However, limited by the capabilities to study and experimentally validate lncRNAs, there is a pressing need for specialized workflows to select functional and biologically relevant lncRNAs from those exclusively detected by Kallisto.

Biologically relevant lncRNAs are uncovered by ELATUS

Now that we have established that a significant number of lncRNAs were exclusively detected when the scRNA-seq data was analyzed with Kallisto, we observed that the majority of them were tissue-specific, reaffirming the well-known specificity of lncRNAs[7] **(Fig. 4a, Supplementary Fig. 8a).** However, the biological relevance of these lncRNAs remains to be experimentally validated. Interestingly, they showed a significant enrichment among lncRNAs identified by CRISPRi screenings in multiple human cell lines[10] (hypergeometric test) (**Supplementary Table 2, Supplementary Fig. 9**), indicating their role in supporting cellular functions. Next, in order to provide a curated list of lncRNAs likely of having biological importance from the thousands of lncRNAs exclusively identified by Kallisto, we implemented a workflow for Elucidating biologically relevant lncRNAs annotated transcripts previously undetected by standard scRNA-seq analysis, termed ELATUS. Using the raw count matrix from Kallisto, we performed a quality control and normalization steps and we reasoned that a biologically important lncRNA should be expressed over a threshold and should have a highly cell type-specific expression pattern (**Fig. 4b**, see Methods). Specifically, after integrating samples originated from the same tissue, ELATUS retained highly-expressed Kallisto-exclusive lncRNAs (i.e., more than 250 counts and present in more than 25 cells according to Kallisto and containing less than 10 counts according to Cell Ranger) and selected those which were 40 times more expressed according to Kallisto than to Cell Ranger and that were highly specific (SI > 0.15). Notably, ELATUS identified 87 cell type-specific and highly expressed lncRNAs, which added to the 173 lncRNAs that are hits in the CRISPRi screenings[10] (**Supplementary Table 3**), defined a set of 260 lncRNAs that are overlooked by Cell Ranger and exhibit characteristics of functional lncRNAs.

Then, to further explore the potential of the ELATUS workflow in identifying lncRNAs with significant biological roles, we decided to investigate 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 Cell Ranger and Kallisto (see Methods). After removing low-quality cells and poorly-expressed genes, we observed an important fraction of highly-expressed lncRNAs exclusively identified by Kallisto (1037 in total) in every sample,

368 whereas most protein-coding genes were commonly detected (**Fig. 4c**),
369 confirming our previous observations.
370
371 Next, we integrated the five TNBC samples, and cells were classified into major
372 and minor cell types using canonical markers (**Fig. 4d-e, Supplementary Fig.**
373 **8b-c, Supplementary Table 1**) and reference datasets[52]. Further, we observed
374 some lncRNAs overlooked by Cell Ranger that had unique expression patterns
375 (**Fig. 4f**). Among them, using ELATUS we identified several cell type-specific
376 candidates, such as *WT1-AS* and *AL133679.1*, which are plasmablasts-specific
377 lncRNAs, or *AC009312.1* that is highly-expressed in dendritic cells (**Fig. 4g**).
378 Interestingly, we also recognized *AL121895.1*, that was specific of breast
379 cancer epithelial cells and could be a potential marker of these tumorigenic
380 cells. These findings remark the importance of ELATUS for uncovering cell type
381 and cancer type-specific lncRNAs in scRNA-seq experiments.
382
383 **ELATUS-identified *AL121895.1* is a *cis*-repressor that participates in**
384 **breast cancer proliferation**
385 Once established that our proposed workflow is able to detect cell type-specific
386 lncRNAs that are missed by Cell Ranger, we next aimed to experimentally
387 validate their expression patterns in different cell types. To do that, we selected
388 *AL121895.1* and *WT1-AS*, which according to ELATUS are specific to breast
389 cancer epithelial cells and plasmablasts, respectively (**Fig. 5a, Supplementary**
390 **Fig. 8d**). We independently analyzed their expression in MDA-MB-231, a
391 human epithelial breast cancer cell line, and in KMS-12-BM, a plasma cell line
392 of multiple myeloma that represents later stages of B-cell differentiation[53–55],
393 similar to plasmablasts. Experimental detection by RT-qPCR confirmed the
394 expression patterns of these lncRNAs that were found functionally relevant by
395 ELATUS, since *AL121895.1* was significantly enriched in breast cancer cells,
396 while *WT1-AS* had a significantly higher expression in multiple myeloma cells
397 (**Fig. 5b**).
398
399 Given that *AL121895.1* is specific to breast cancer epithelial cells, we next
400 investigated its function in these tumoral cells by knocking it down using siRNAs
401 (**Fig. 5c, Supplementary Fig. 8d**). Interestingly, we observed that the
402 knockdown of *AL121895.1* significantly decreased the proliferation rate of
403 breast cancer cells in an MTS proliferation assay (**Fig. 5d**), indicating a role in
404 these tumorigenic cells. Furthermore, we noticed that its knockdown increased
405 the expression of the antisense protein-coding gene, *EPB41L1* (**Fig. 5e,**
406 **Supplementary Fig. 8e**), suggesting a previously undocumented *cis*-repressor
407 function for *AL121895.1*. Inspection of the expression of *AL121895.1* and
408 *EPB41L1* in the scRNA-seq of the TNBC samples, also confirmed a strong
409 anticorrelation between them in each cell type (**Fig. 5f**). Further, since lncRNAs
410 with similar k-mer content are functionally related[49], we clustered, by K-means,
411 the k-mer content of *AL121895.1* together with the k-mer composition of
412 previously described lncRNAs that are either *cis*-activators or *cis*-repressors[49].
413 The clustering grouped *AL121895.1* together with six other lncRNAs, from

414 which five were proved *cis*-repressors (**Fig. 5g**), supporting its repressive role
415 on its neighboring gene.
416
417 Together, these results demonstrate the biological significance of lncRNAs
418 missed by Cell Ranger and highlight the need for ELATUS, the presented
419 optimized scRNA-seq workflow in order to unlock cellular features encoded by
420 lncRNAs.
421
422
423 **Discussion**
424 It has been widely established that lncRNAs regulate gene expression by
425 different mechanisms at multiple levels[5,6,11]. However, given their high cell type
426 specificity[7,8,10], the current knowledge of lncRNAs can be greatly expanded with
427 the application of single-cell technologies. Indeed, previous studies have
428 showed that lncRNAs alone can identify cell types[1,50] and that there are certain
429 cell groups that can only be distinguished by them[3]. Nevertheless, compared to
430 protein-coding genes, the annotation of lncRNAs is less well established and
431 more prone to inaccuracies[15], which complicates its correct quantification. In
432 fact, as previously probed in bulk RNA-seq experiments[39], pseudoaligners are
433 promising alternatives to the closed-source and aligner-based Cell Ranger,
434 which is the standard scRNA-seq preprocessing pipeline. In this work we have
435 shown, in a wide and diverse set of public scRNA-seq datasets, that the
436 detection and quantification of lncRNAs is severely affected by the
437 preprocessing choice and the presented workflow tailored to lncRNAs,
438 ELATUS, is essential for identifying functional lncRNAs previously undetected.
439
440 ELATUS requires the sequencing data to be preprocessed by Kallisto, due to its
441 increased detection of lncRNAs. Although previous studies have already
442 described that Kallisto is able to detect a higher number of lncRNAs, they did
443 not delve into the potential origin of this behavior. They either note that Kallisto
444 provided a comparable increment in the detection of both lncRNAs and protein-
445 coding genes[36], or attributed it to poorly expressed lncRNAs solely captured
446 with Kallisto that did not provide a significant biological signal gain[27]. We
447 hypothesize that this increased detection is not happening with Salmon due to
448 its selective alignment strategy, in which the authors recommend including a set
449 of decoy sequences to provide a more precise mapping[56]. Further, preceding
450 investigations have alerted that there is a reduction in quantification accuracy of
451 pseudoaligners due to the spurious mapping of intronic reads[29,42], which can
452 lead to false expression of particular genes[36]. The developers of Kallisto noted
453 this phenomenon to be plausible but rare[32]. Given the pressing need to better
454 characterize lncRNAs at single-cell resolution, we aimed to understand the
455 reasons underlying its increased detection by Kallisto. Therefore, we have
456 performed a detailed characterization of the lncRNAs exclusively detected with
457 Kallisto compared to those commonly captured. We have found that the
458 explanation that spurious alignments are caused by poor quality mapping of
459 low-complexity sequences does not, at least entirely, apply to lncRNAs. Both

460 groups of lncRNAs are comparable and there is not a particular feature that
461 causes the extra-detection. However, the fact that lncRNAs exclusively detected
462 by Kallisto have a significantly higher number of exons supports the application
463 of pseudoaligners and provides an alternative to the spurious mapping that also
464 explains its increased detection rate.
465
466 To inspect the advice against pseudoaligners, we have both computationally
467 and experimentally validated the expression patterns obtained by Kallisto
468 suggesting that its increased detection is, at least in part, due to the
469 identification of readily expressed functional transcripts[57]. In fact, lncRNAs only
470 detected by Kallisto agree better with ATAC-seq data than those detected by
471 Cell Ranger, indicating that the pseudolignment is detecting bona fide
472 transcripts generated from regions of open chromatin. Moreover, to inspect the
473 thousands of lncRNAs exclusively detected with Kallisto, we implemented a
474 computational workflow, ELATUS (available online as an R package), that
475 uncovers highly-specific and biologically important lncRNAs, while minimizing
476 the risk of false expression of the exclusively detected genes provided by
477 Kallisto, as previous investigations have warned[29,42]. In fact, ELUTUS provided
478 a list of plausible cell type-specific and functionally relevant lncRNAs that were
479 missed by Cell Ranger. It is of particular relevance the previously
480 uncharacterized lncRNA *AL121895.1*, which we have found is a *cis*-repressor
481 specific of breast cancer epithelial cells and whose functionality at single-cell
482 level could not be determined by Cell Ranger. Indeed, the knockdown of
483 *A121895.1* decreased the proliferation of breast cancer cells. *AL121895.1*
484 regulates *EPB41L1* expression, encoding a multifunctional protein that
485 mediates interactions between the erythrocyte cytoskeleton and the overlying
486 plasma membrane, although it is also expressed in other tissues[58,59]. Our
487 results indicate that the regulation of *EPB41L1* by *AL121895.1* could represent
488 a feature of breast tumor cells and a possible cancer vulnerability that could
489 only be uncovered by the proposed workflow.
490
491 The presented data evidences the potential impact of ELATUS to unveil
492 important biological roles of lncRNAs and to expand the map of interactions, in
493 individual cell populations, between the expression of a previously undetected
494 by Cell Ranger lncRNA and the nearby protein-coding gene. Moreover, it
495 exemplifies how scRNA-seq can inform outstanding mechanistic questions,
496 such as the cis vs trans regulatory roles of lncRNAs.
497
498 It should be noted that the library preparation method could also influence the
499 detection of lncRNAs. Conventional 10x Genomics scRNA-seq library
500 preparation protocols target only polyA transcripts. Since an important fraction
501 of lncRNAs are not polyadenylated[6], these technologies cannot achieve a
502 complete map of the transcriptome. Therefore, alternative scRNA-seq
503 preparation protocols that capture both polyadenylated and non-polyadenylated
504 transcripts, processed following the ELATUS workflow could potentially reveal

505 an expanded number of functional non-coding transcripts that participate in
506 important cellular functions.
507
508 Another direction that should be explored is the influence of the reference
509 annotation. In this sense, the application of an intronic reference, already
510 recommended for snRNA-seq and that considers both mature and unmature
511 RNAs has been widely discussed in the scRNA-seq community. While some
512 groups recommended its inclusion to avoid spurious mapping of intronic
513 reads[42], other studies have concluded that their addition produces more noise
514 and is not translated into a significant biological gain[27]. Also, it poses a
515 disjunctive when an exon of a gene is overlapped by an intron of another gene
516 since every read falling in that region would be considered ambiguous
517 (**Supplementary Fig. 10a-b**). There is no clear consensus on how to resolve
518 these situations. For example, Cell Ranger, which in their last version (v.7)
519 counts introns by default, arbitrarily assigns those reads to the mature exons,
520 while the developers of Kallisto and Salmon[60,61] have recently suggested
521 alternative strategies to determine its origin based on their k-mer content and
522 flanking regions. Recently, other studies have recommended the combination of
523 an intronic reference together with a corrected and adjusted annotation of 3'
524 UTR regions[62] while others claim that a *de novo* transcriptome assembly
525 approach provides a more precise annotation of lncRNAs[50]. Future investigation
526 combining an improved reference annotation together with ELATUS could
527 provide significant improvements in gene detection, especially for less studied
528 biotypes such as lncRNAs.
529
530 Finally, in the proposed workflow, we reasoned that a functional lncRNAs
531 should be highly specific. While, this is true regarding the defined properties of
532 lncRNAs[7–10], the existance of several ubiquitous lncRNAs playing essential
533 roles[63,64] should be noted. Here, we propose an optimized computational
534 workflow for analyzing scRNA-seq experiments that has the potential to unlock
535 novel cellular features and transcriptional complexity, increasing the insights
536 into cell identity and lncRNA biology.
537
538
539 **Methods**
540 **Single-cell RNA-seq preprocessing worflows**
541 The following scRNA-seq preprocessing pipelines were benchmarked: Cell
542 Ranger, STARsolo, Kallisto-Bustools (referred as Kallisto) and Salmon-Alevin
543 (referred as Salmon). All pipelines were executed with the default
544 recommended parameters in the user guides. For scRNA-seq analysis, Cell
545 Ranger count was run in version 3.0.1, STAR in version 2.7.9, Kallisto in
546 version 0.46.1, Bustools in version 0.40.0 and Salmon in version 1.4.0. For the
547 pulmonary fibrosis dataset, since it was necessary to split cDNA sequence in
548 more than one file, Kallisto was run in version 0.46.2 following authors
549 indications[65].
550

551 **Reference annotation and generation of the indexes**
552 Human and mouse reference genome, transcriptome and annotation were
553 downloaded from GENCODE[13]. Particularly, for human, hg38 (v37), and for
554 mouse, mm10 (v27), were selected. We created the indexes for each
555 preprocessing pipeline following recommended settings. Specifically, Cell
556 Ranger and STARsolo were indexed against the entire genome, Kallisto against
557 the transcriptome and Salmon was indexed to perform selective alignment to
558 the transcriptome with full decoys as suggested by both the authors and
559 independent benchmark studies[29,36,42,66]. The commands for preprocessing and
560 generating the indexes for each preprocessing tool can be found in
561 https://github.com/kikegoni/manuscript_scRNAseq_lncRNAs.
562
563 To analyze the evolution in the number of annotated protein-coding genes and
564 lncRNAs, we analyzed the annotation provided by GENCODE since version 7,
565 published in 2010, until the last version (v.44), published in December 2022.
566
567 **scRNA-seq public datasets**
568 Different scRNA-seq prepared with 10x Genomics protocols, with both v2 and
569 v3 chemistries, were used. Most of them were public, while sequencing data
570 regarding TNBC samples were manually prepared. The description of each
571 dataset as well as their link to access the sequencing data is provided in
572 **Supplementary Table 4**.
573
574 **scRNA-seq quality control, gene detection and post-processing steps**
575 Raw count matrices were used to standardize preprocessing pipelines as input
576 for quality control, where we followed common scRNA-seq computational
577 guidelines[67]. Specifically, emptydrops[68] was applied to distinguish empty
578 droplets from cells in each dataset processed with Cell Ranger, STARsolo and
579 Kallisto. On the other side, Salmon, employs a whitelisting filtering strategy to
580 filter empty droplets and it does not output the raw count data. To account for
581 that and standardize filtering strategies, the minimum number of counts
582 surviving emptydrops filtering in Cell Ranger, STARsolo and Kallisto was
583 selected as an additional threshold to filter cells in Salmon with fewer counts
584 than this defined threshold. Potential doublets were then identified and removed
585 with scDblFinder[69]. Finally, cells with high mitochondrial content and an
586 abnormally high number of counts were also filtered.
587
588 Once low-quality cells have been removed, in order to compare the detection
589 and quantification of protein-coding genes and lncRNAs, poorly-expressed
590 genes were also filtered by applying a gradient of thresholds on the expression.
591 The different thresholds applied retained those genes with more than 1) 250
592 counts and present in more than 25 cells, 2) 100 counts and present in more
593 than 10 cells, 3) 50 counts and present in more than 5 cells and 4) 25 counts
594 and present in more than 3 cells. These thresholds were also applied in the
595 characterization of the genes exclusively identified by Kallisto compared to
596 genes commonly found by all pipelines. Due to the differences in the preserved

597 number of cells in each scRNA-seq dataset, both absolute numbers and
598 percentages have been used to represent the differences and the overlap of the
599 detected highly-expressed genes by Kallisto and Cell Ranger.
600
601 After the quality control was completed, normalization was performed using
602 logNormCounts function from scuttle R package[70]. Dimensionality reduction
603 was conducted using runPCA, runTSNE and runUMAP functions from scater R
604 package[71]. Next, clustering was performed on a generated shared nearest-
605 neighbour (SNN)[72] graph using the Louvain community detection algorithm to
606 cluster the cells[73,74]. These clusters were manually annotated to cell types using
607 canonical markers (**Supplementary Table 1**). Subtypes in TNBC integrated
608 data were distinguished using the annotation program JIND[75] selecting as a
609 reference a dataset consisting of ~45000 cells from TNBC samples[52]. We
610 assessed the correlation (Spearman) between the normalized expression of
611 *AL121895.1* and *EPBL41L1* per these cellular subtypes in each cell that
612 expressed either *AL121905.1* or *EPBL41L1* (or both).
613
614 **Single-cell multiome analysis**
615 In 3k PBMCs sequenced with single cell multiome, scATAC-seq raw data has
616 been directly downloaded from 10X Genomics website[48] on which scATAC-seq
617 specific thresholds were first applied to remove low-quality nuclei. In particular,
618 nuclei with very few or excessive ATAC-seq counts were filtered, as well as
619 those with high nucleosome signal or little enrichment at the TSS[76].
620
621 Gene expression data was obtained with Cell Ranger count (version 5.0.1)
622 using --include-introns option. Regarding Kallisto (version 0.46.1) the index was
623 generated to include both introns and exons using --workflow lamanno
624 parameter.
625
626 For the raw RNA-seq matrices, those nuclei the fit the scATAC-seq quality
627 control thresholds in both Cell Ranger and Kallisto were retained. Nuclei with
628 very few RNA-seq counts or very high mitochondrial content were further
629 removed. Next, ATAC-seq data from these high-quality nuclei were normalized
630 using a Latent Semantic Indexing approach. “Weighted nearest neighbour”
631 (WNN) analysis was then performed to integrate the ATAC-seq with the gene
632 expression obtained by Cell Ranger and Kallisto. This integrated data was used
633 for generating the clusters (following indications in scRNA-seq post-processing
634 section) that were manually assigned to different cell types according to the
635 expression of established marker genes.
636
637 To compare the similarity between the ATAC-seq signal and the RNA-seq gene
638 expression, the GeneActivity function from Signac R package[77] was applied to
639 obtain a gene activity matrix by counting the scATAC-seq fragments that fall in
640 each gene body (+2Kb upstream from the TSS). Then, for every high-quality
641 nucleus we compared (student’s t-test) the number of genes that have
642 simultaneous ATAC-seq-signal and RNA-seq expression when preprocessed

643 with Cell Ranger or with Kallisto. A gene was defined to be simultaneously
644 activated if its ATAC-seq signal and its RNA-seq expression were higher than a
645 gradient of defined thresholds. Specifically, from less to more restrictive
646 thresholds; if they had at least 1) 1 read in ATAC-seq and 1 UMI in RNA-seq, 2)
647 3 reads in ATAC-seq and 3 UMIs in RNA-seq, 3) 6 reads in ATAC-seq and 6
648 UMIs in RNA-seq, 4) 6 reads in ATAC-seq and 11 UMIs in RNA-seq and 5) 11
649 reads in ATAC-seq and 11 UMIs in RNA-seq.
650
651 The differences, per cell, were also represented as an odds ratio (in log2 scale)
652 showing the likelihood of having more genes simultaneously activated with
653 Kallisto than with Cell Ranger. Further, the ratio (in log2 scale), of the number of
654 nuclei for which there were more simultaneous activation when the snRNA-seq
655 was processed with Kallisto than with Cell Ranger has been computed. A
656 positive ratio indicates a better correspondence between ATAC-seq and Kallisto
657 than between ATAC-seq and Cell Ranger, while a negative ratio indicates the
658 opposite.
659
660 **Exclusive vs common genes: Length, number of exons, repeat content**
661 **and k-mer analysis**
662 To investigate the length, the number of exons and repeat content of both
663 exclusive and commonly detected genes in every scRNA-seq dataset, the
664 longest isoform of each gene was selected. The annotation of repeats for both
665 human and mouse genomes was downloaded from RepeatMasker (version
666 4.1.5)[78], where we considered all distinct types of repeats with the exception of
667 microsatellites repeats. We calculated the ratio between the percentage of the
668 sequence of exclusive features covered by repeats divided by the percentage of
669 the sequence of common features covered by repeats.
670
671 We compared the k-mer content of all lncRNAs using SEEKR[49], a software
672 developed for sequence evaluation through k-mer content, where we calculated
673 the 6-mer functionally-related communities using the canonical isoform of each
674 gene following the default recommended settings by the authors[79]. For the K-
675 means (stats R package) clustering of *AL121895.1* according its k-mer content
676 and the k-mer content of previously described cis-activators or cis-repressors[49],
677 we considered the isoform *AL121895.1-EST00000441208*. This is the isoform
678 to which Kallisto assigned most of scRNA-seq reads and for which qPCR
679 primers and siRNAs were designed.
680
681 **Exclusive vs common genes: Specificity index**
682 In order to implement the Specificity Index (SI), each scRNA-seq datasets was
683 clustered across distinct ranges of number and sizes of clusters, from fewer and
684 bigger clusters to more but smaller clusters. We started by splitting the cells in
685 5-9 clusters, then 10-15 clusters, then 16-22 clusters and finally generating a
686 very detailed subdivision with at least 23 clusters. The SI metric was then
687 designed in order to inform if a lncRNA is more specific of big or small
688 subpopulations. We implemented the SI following the Shannon-Entropy

689 specificity (HS) formulation defined in TSPEX, a library with several specificity
690 metrics[51]. So, in order to define the SI for each gene *g*, we first calculated its
691 mean expression x_i in each cluster, i = 1,2…n, where n is the number of
692 clusters. Next, we calculated for each gene, the proportion of mean expression
693 in each cluster, P_i :

694 $$P_i = \frac{x_i}{\sum_{i=1}^{n}(x_i)}$$

695

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

700 $$SI = 1 + \sum_{i=1}^{n} P_i * log_n P_i$$

701

702 The SI is ranked from [0-1], from genes very ubiquitously expressed to very
703 cluster-specific genes. A gene whose expression is equally distributed across
704 different clusters will have a SI of 0, while if a gene is exclusively expressed in
705 one cluster its SI will be 1.
706

707 **ELATUS workflow to define a list of biologically relevant lncRNAs**
708 **overlooked by Cell Ranger**

709 On the one side, to get the list of biologically relevant lncRNAs, we analyzed the
710 288 CRISPR functionally validated lncRNAs in multiple human cell lines[10]. We
711 tested if their ovelap with those lncRNAs exclusively identified by Kallisto, in
712 every human scRNA-seq dataset, was significant using a hypergeometric test.
713 In total, there was an overlap of 173 Kallisto-exclusive lncRNAs.
714

715 On the other side, we implemented ELATUS in order to inspect the highly-
716 expressed lncRNAs exclusively detected with Kallisto and to assess their
717 biological relevance. Therefore, we started by importing the raw count matrices
718 obtained with both Cell Ranger and Kallisto and we performed emptydropts to
719 distinguish empty droplets from cells. Next, we removed potential multiplets and
720 performed a quality-control filtering, followed by a normalization and clustering
721 steps. Then, those lncRNAs with more than 250 counts and present in more
722 than 25 cells by Kallisto were selected and we calculated their Specificity Index
723 (SI) as previously indicated (Specificity Index section). From them, we retained
724 those lncRNAs for which Cell Ranger assigned less than 10 counts, that were
725 40 times more expressed according to Kallisto than to Cell Ranger and that,
726 according to Kallisto, had a SI > 0.15. To include a representative set of cluster
727 sizes to calculate the SI, scRNA-seq datasets were divided in different cluster
728 sizes, from 10 to 19 clusters (**Supplementary Table 4**). ELATUS, which is
729 openly available as an R package in https://github.com/kikegoni/ELATUS
730 identified a list of the 87 lncRNAs, that added to the 173 lncRNAs that are hits in
731 the CRISPRi screenigs, defined a list of 260 Kallisto-exclusive lncRNAs that
732 exhibit characteristics of functional lncRNAs.
733

734 **Statistical analysis and data plotting**

735 Post-processing analysis were performed in R (version 4.1.2). Barplots and
736 violin plots were represented with ggplot2 (v.3.4.2), where ggpubr (v.0.4.0) was
737 used to test statistical test significance. The specific statistical test for each
738 analysis is detailed in its figure caption, where *ns* represents p-value > 0.1, *
739 represents p-value <= 0.1, ** represents p-value <= 0.05, *** represents p-value
740 <= 0.005 and **** represents p-value <= 0.0005.
741
742 From Seurat (v.4.0.1)[80], DimPlot function was used to plot UMAP dimensionality
743 reduction plots and DotPlot and FeaturePlots functions were applied to evaluate
744 gene expression in different cell types and dimensionality reduction spaces,
745 respectively. UpSet plots were generated with UpSetR (v.1.4.0)[81] and ggupset
746 (v.0.3.0). In the analysis of single-cell multiome data, Signac R package (v.
747 1.9.0)[77] was applied to create the coverage and expression plots.
748
749 **Cell lines and growth conditions**
750 MDA-MB-231 cells were cultured in DMEM (GIBCO), supplemented with 10%
751 fetal bovine serum (GIBCO) and 1x penicillin/streptomycin (Lonza), while KMS-
752 12 cells were grown in RPMI-1640 (GIBCO) medium with 20% fetal bovine
753 serum (GIBCO), 1% penicillin, and 2% Hepes. All of them were maintained at
754 37˚C and 5% CO_2.
755
756 **RNAi**
757 For RNA knockdown, siRNAs, which were designed using the i-Score designer
758 tool and purchased from Sigma (**Table 1**), were transfected for 24 hours at 40
759 nM final concentration. MDA-MB-231 cells were transfected with Lipofectamine
760 2000 (Invitrogen) in Serum-free Opti-MEM (GIBCO), following manufacturer
761 instructions.
762

Directed transcript	siRNA sequence
AL121895.1-ENST00000441208.1 siRNA1	GGGCAGAGAUUCAACCAAA
AL121895.1-ENST00000441208.1 siRNA2	CAAGUAUUUUGAUUCCAAA
Scramble siRNA	CAGUCGCGUUUGCGACUGG

763 **Table 1**. List and sequence of siRNAs designed to knockdown lncRNAs.
764
765 **Proliferation assay**
766 Cell proliferation was measured using the CellTiter96 Aqueous Non-Radioactive
767 Cell Proliferation Assay (MTS) kit (Promega). After 24 hours transfection, 1000
768 MDA cells were cultured in M-96 plate wells. Every 24 hours, 20 µL of MTS
769 reagent (Promega) were added to culture media, and incubated for 2 hours,
770 prior to 490 λ measurement. Triplicate measures were normalized to day 0, and
771 statistical differences between control and experimental conditions at day 3
772 were calculated with a two-tailed student t-test.
773

774 **RNA extraction, processing, and RT-qPCR**
775 Cell preparations were fixed with TRIzol (Sigma), and RNA precipitated with
776 isopropanol. RNA extraction was followed by Turbo DNAse (Invitrogen)
777 digestion for 30 minutes at 37ºC. For RT-qPCR, 1 µg RNA was reverse-
778 transcribed using the High-Capacity cDNA Reverse Transcription Kit (Applied
779 30 Biosystem) with random hexamer primers, following manufacturer
780 instructions. The obtained cDNA was analysed by quantitative PCR (qPCR)
781 using iTaq Universal SYBR Green supermix (Bio-Rad) in a ViiA™ 7 Real-Time
782 PCR System machine (ThermoFisher), all reactions performed in quadruplicate.
783 GAPDH RNA levels were used for normalization. Statistical differences between
784 relative RNA levels were calculated by unpaired two-tailed Student's *t*-test. RT-
785 qPCR primers were self-designed or designed with the NCBI Primer designing
786 tool, and purchased from Metabion (**Table 2**).
787

Oligonucleotide name	Sequence
WT1-AS #1 Forward	TGGGAGCCTCGTCTCACT
WT1-AS #1 Reverse	GTCGGAGAAGTCCAAGTCGG
WT1-AS #2 Forward	CCCGGAACCGTTAGGTACTG
WT1-AS #2 Reverse	GGTGGCATCAGAACAGACGA
AL121895.1 #1 Forward	GTATGGCCCCCATAGGCATC
AL121895.1 #1 Reverse	CCCTGCCTGGGGAATGAAAA
AL121895.1#2 Forward	CAGCAGCGGATGAAGAGCA
AL121895.1 #2 Reverse	GAAGGCAGGCAGTCAAGAAT
EPB41L1 #1 Forward	ACTGTGGAAGGTCTGCATCG
EPB41L1 #1 Reverse	GAACTTGGAGCCCATCACCA
EPB41L1 #2 Forward	AGTTCTACCCGCCTGATCCT
EPB41L1 #2 Reverse	AGTAGGGCATGCGTGACAAA

788 **Table 2**. List and sequence of oligonucleotides designed for RT-qPCR.
789
790 **Human samples**
791 All patients participating in the study provided written informed consent. The
792 study and the use of all clinical materials have been approved by the Research
793 Ethics Committee of the University of Navarra under decision number
794 2021.058Mod1. Samples and data from patients included in the study were
795 provided by the Biobank of the University of Navarra and were processed
796 following standard operating procedures approved by the Ethical and Scientific
797 Committees.
798
799 **TNBC sample preparation and preparation of the scRNA-seq libraries**
800 Five biopsies of tissue from different patients of breast cancer were processed
801 following manufacturer's instructions of Human Tumor Dissociation Kit from
802 Miltenyi Biotec. Briefly, 2-3 tissue cylinders per patient were cut into small slices
803 of 2–4 mm, digested with enzymes H, R and A, and dissociated with
804 gentleMACS Dissociator (Miltelnyi Biotec). After 30 minutes incubation at 37ºC
805 and a short centrifugation step (1800 rpm, 1 minute at 4ºC), sample material at

806 the bottom of the tube was collected. cell suspension was then moved to a
807 Falcon cell strainer (70 µm) placed on a 50 mL tube, washed with 20 mL of
808 DMEM, and centrifuged at 1800 rpm for 5 minutes at 4ºC. Pelleted cells were
809 washed with 1ml of PBS 1X 0.05%BSA, supplemented with 5ul RNAse OUT
810 (Invitrogen), and transferred to a Dolphin tube (Sorenson) where cell
811 suspension was centrifuged at 1800 rpm for 5 minutes at room temperature.
812 Viability (>70%) of resuspended cells was corroborated with cellometer
813 (Nexcelom).
814
815 The transcriptomes of 16,000-20,000 cells were examined using Single Cell 3'
816 Reagent Kits v3.1 (10X Genomics) according to the manufacturer's instructions.
817 Briefly, 17000 to 20000 cells were loaded at a concentration of 1000 cells/µL on
818 a Chromium Controller instrument (10X Genomics) to capture single cells in gel
819 bead-in-emulsions (GEMs). In this step, each cell was encapsulated with
820 primers containing a fixed Illumina Read 1 sequence, a cell-identifying 16 bp
821 10x Genomics barcode, a 12 bp Unique Molecular Identifier (UMI) and a poly-
822 dT sequence. Upon cell lysis, reverse transcription yielded full-length, barcoded
823 cDNA. This cDNA was then released from the GEMs, PCR-amplified and
824 purified with magnetic beads (SPRIselect, Beckman Coulter). Enzymatic
825 Fragmentation and Size Selection was used to optimize cDNA size prior to
826 library construction. Fragmented cDNA was then end-repaired, A-tailed and
827 ligated to Illumina adaptors. A final PCR-amplification with barcoded primers
828 allowed sample indexing. Library quality control and quantification was
829 performed using Qubit 3.0 Fluorometer (Life Technologies) and Agilent's 4200
830 TapeStation System (Agilent), respectively. Sequencing was performed in a
831 NextSeq2000 (Illumina) (Read1: 28; Read2: 91; i7 index: 8) at an average
832 depth of 500000 reads/sample.
833
834
835 **Data availability**
836 TNBC scRNA-seq data has been publicly stored in NCBI with the following
837 identifier: GSE246142.
838
839
840 **Code availability**
841 The code and scripts to reproduce the results is freely available and can be
842 accessed in https://github.com/kikegoni/manuscript_scRNAseq_lncRNAs.
843 ELATUS R package for elucidating biologically relevant lncRNA annotated
844 transcripts previously undetected using scRNA-seq is available at
845 https://github.com/kikegoni/ELATUS.
846
847
848
849
850
851

852 **References**
853

854 1. Rahman, R. U., Ahmad, I., Sparks, R., Saad, A. Ben & Mullen, A.
855 Singletrome: A method to analyze and enhance the transcriptome with
856 long noncoding RNAs for single cell analysis.
857 doi:10.1101/2022.10.31.514182.
858 2. Luo, H. et al. Single-cell Long Non-coding RNA Landscape of T Cells in
859 Human Cancer Immunity. Genomics Proteomics Bioinformatics 19, 377–
860 393 (2021).
861 3. Zheng, L. L. et al. ColorCells: a database of expression, classification and
862 functions of lncRNAs in single cells. Brief Bioinform 22, 1–11 (2021).
863 4. Santus, L. et al. Single-cell profiling of lncRNA expression during Ebola
864 virus infection in rhesus macaques. Nature Communications 2023 14:1
865 14, 1–14 (2023).
866 5. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long
867 non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol
868 doi:10.1038/s41580-020-00315-9.
869 6. Mattick, J. S. et al. Long non-coding RNAs: definitions, functions,
870 challenges and recommendations. Nature Reviews Molecular Cell Biology
871 2023 24:6 24, 430–447 (2023).
872 7. Cabili, M. et al. Integrative annotation of human large intergenic
873 noncoding RNAs reveals global properties and specific subclasses.
874 Genes Dev 25, 1915–1927 (2011).
875 8. Liu, S. J. et al. Single-cell analysis of long non-coding RNAs in the
876 developing human neocortex. (2016) doi:10.1186/s13059-016-0932-1.
877 9. Atanasovska, B. et al. A liver-specific long noncoding RNA with a role in
878 cell viability is elevated in human nonalcoholic steatohepatitis. Hepatology
879 66, 794–808 (2017).
880 10. Liu, S. J. et al. CRISPRi-based genome-scale identification of functional
881 long noncoding RNA loci in human cells. Science (1979) 355, (2017).
882 11. Huarte, M. The emerging role of lncRNAs in cancer. Nature Medicine
883 2015 21:11 21, 1253–1261 (2015).
884 12. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human
885 transcriptome. Nature Genetics 2015 47:3 47, 199–208 (2015).
886 13. Frankish, A. et al. GENCODE: reference annotation for the human and
887 mouse genomes in 2023. Nucleic Acids Res 51, D942–D949 (2023).
888 14. Hezroni, H. et al. Principles of Long Noncoding RNA Evolution Derived
889 from Direct Comparison of Transcriptomes in 17 Species. Cell Rep 11,
890 1110–1122 (2015).
891 15. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. &
892 Johnson, R. Towards a complete map of the human long non-coding RNA
893 transcriptome. Nature Reviews Genetics 2018 19:9 19, 535–548 (2018).
894 16. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age.
895 Nature Communications 2020 11:1 11, 1–4 (2020).
896 17. Zeisel, A. et al. Molecular Architecture of the Mouse Nervous System.
897 Cell 174, 999-1014.e22 (2018).
898 18. Prescott, S. L., Umans, B. D., Williams, E. K., Brust, R. D. & Liberles, S.
899 D. An Airway Protection Program Revealed by Sweeping Genetic Control
900 of Vagal Afferents. Cell 181, 574-589.e14 (2020).

19. La Manno, G. et al. RNA velocity of single cells. Nature 2018 560:7719 560, 494–498 (2018).
20. Macosko, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214 (2015).
21. Klein, A. M. et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 161, 1187–1201 (2015).
22. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications 2017 8:1 8, 1–12 (2017).
23. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols 2018 13:4 13, 599–604 (2018).
24. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nature Reviews Immunology 2017 18:1 18, 35–45 (2017).
25. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 53, 1005–1005 (2021).
26. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 15, e8746 (2019).
27. You, Y. et al. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. Genome Biol 22, 339 (2021).
28. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
29. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. bioRxiv 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.
30. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 2016 34:5 34, 525–527 (2016).
31. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 2017 14:4 14, 417–419 (2017).
32. Melsted, P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nature Biotechnology 2021 39:7 39, 813–818 (2021).
33. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. Bioinformatics 35, 4472–4473 (2019).
34. 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).
35. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nature Communications 2019 10:1 10, 1–11 (2019).
36. 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, (2022).
37. Du, Y., Huang, Q., Arisdakessian, C. & Garmire, L. X. Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment. G3 Genes|Genomes|Genetics 10, 1775–1783 (2020).

38. Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. doi:10.1101/gr.132159.111.
39. 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).
40. 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.
41. 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.
42. He, D. et al. Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. Nature Methods 2022 19:3 19, 316–322 (2022).
43. 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).
44. Schupp, J. C. et al. Integrated Single-Cell Atlas of Endothelial Cells of the Human Lung. Circulation 144, 286–302 (2021).
45. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv 6, (2020).
46. 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.
47. 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.
48. 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.
49. Kirk, J. M. et al. Functional classification of long non-coding RNAs by k-mer content. Nature Genetics 2018 50:10 50, 1474–1482 (2018).
50. Bitar, M. et al. Redefining normal breast cell populations using long noncoding RNAs. Nucleic Acids Res (2023) doi:10.1093/nar/gkad339.
51. 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 (2020) doi:10.21203/RS.3.RS-51998/V1.
52. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. Nature Genetics 2021 53:9 53, 1334–1347 (2021).
53. Namba, M. et al. Establishment of five human myeloma cell lines. In Vitro Cellular & Developmental Biology 25, 723–729 (1989).
54. Edwards, J. C. W. & Cambridge, G. B-cell targeting in rheumatoid arthritis and other autoimmune diseases. Nature Reviews Immunology 2006 6:5 6, 394–403 (2006).

55. 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).
56. Srivastava, A. et al. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol 21, 1–29 (2020).
57. Shainer, I. & Stemmer, M. Choice of pre-processing pipeline influences clustering quality of scRNA-seq datasets. BMC Genomics 22, (2021).
58. Wang, H. et al. Selective effects of protein 4.1N deficiency on neuroendocrine and reproductive systems. Scientific Reports 2020 10:1 10, 1–14 (2020).
59. 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).
60. Hjörleifsson, K. E., Sullivan, D. K., Holley, G., Melsted, P. & Pachter, L. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. doi:10.1101/2022.12.02.518832.
61. He, D., Soneson, C. & Patro, R. Understanding and evaluating ambiguity in single-cell and single-nucleus RNA-sequencing. bioRxiv 2023.01.04.522742 (2023) doi:10.1101/2023.01.04.522742.
62. Pool, A. H., Poldsam, H., Chen, S., Thomson, M. & Oka, Y. Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references. Nature Methods 2023 20:10 20, 1506–1515 (2023).
63. Chakraborty, S. et al. Harnessing the tissue and plasma lncRNA-peptidome to discover peptide-based cancer biomarkers. Scientific Reports 2019 9:1 9, 1–17 (2019).
64. Goyal, B. et al. Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer. BBA-Reviews on Cancer 1875, 188502 (2021).
65. SC5P-R2 sequencing · Issue #226 · pachterlab/kallisto. https://github.com/pachterlab/kallisto/issues/226.
66. Selective Alignment. https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/.
67. Amezquita, R. A. et al. Orchestrating single-cell analysis with Bioconductor. Nature Methods 2019 17:2 17, 137–145 (2019).
68. 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).
69. Germain, P. L., Lun, A., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. F1000Research 2021 10:979 10, 979 (2021).
70. LTLA/scuttle: Clone of the Bioconductor repository for the scuttle package. https://github.com/LTLA/scuttle/.
71. 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).
72. Lun, A. T. et al. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research 2016 5:2122 5, 2122 (2016).
73. Network Analysis and Visualization [R package igraph version 1.5.1]. (2023).

74. igraph – Network analysis software. https://igraph.org/.
75. Goyal, M. et al. JIND: joint integration and discrimination for automated single-cell annotation. Bioinformatics 38, 2488–2495 (2022).
76. Joint RNA and ATAC analysis: 10x multiomic • Signac. https://stuartlab.org/signac/articles/pbmc_multiomic.
77. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. Nature Methods 2021 18:11 18, 1333–1341 (2021).
78. RepeatMasker Home Page. https://www.repeatmasker.org/.
79. CalabreseLab/seekr: A library for counting small kmer frequencies in nucleotide sequences. https://github.com/CalabreseLab/seekr.
80. Hao, Y. et al. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29 (2021).
81. 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).

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.: 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 analysis. A.M. performed experimental validations. 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.

Disclosure of conflicts of interest

The authors declare no competing interests.

Figures

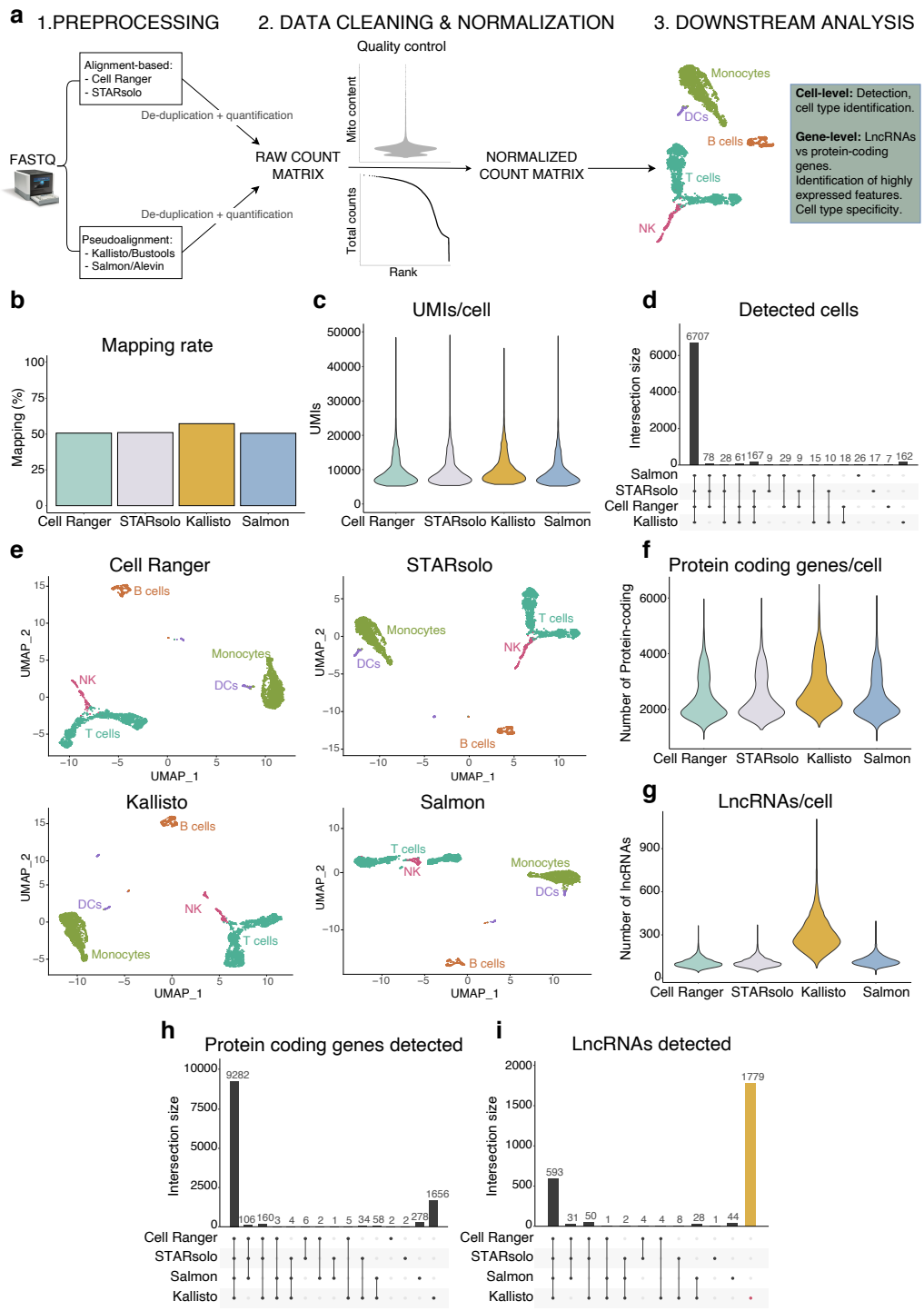

Figure 1. Preprocessing choices strongly affect lncRNA detection in a scRNA-seq dataset consisting of 10k human PBMCs from a healthy donor. a, Detailed step-by-step explanation of the benchmark. First, public fastq files were directly processed with the aligner-based Cell Ranger and STARsolo, as well as with the pseudoaligners Kallisto and Salmon to generate the raw count

data. Next a quality control step was performed to filter empty droplets, cells with high mitochondrial content and potential multiplets, followed by a normalization step. Then, the downstream analysis began with a dimensionality reduction followed by a clustering and cell annotation steps. The detection and quantification of high-quality cells, main cell types, protein-coding genes, lncRNAs and cell type-specific features was compared depending on the preprocessing choice. **b,** Mapping rate by each pipeline. **c,** Number of UMI counts per cell by each pipeline. **d,** UpSet plot showing the overlap of retained high-quality cells by each pipeline. **e,** UMAP plots displaying the populations of 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 (more than 250 counts and present in more than 25 cells) per pipeline. **i,** UpSet plot displaying the overlap of highly-expressed lncRNAs (more than 250 counts and present in more than 25 cells) per pipeline.

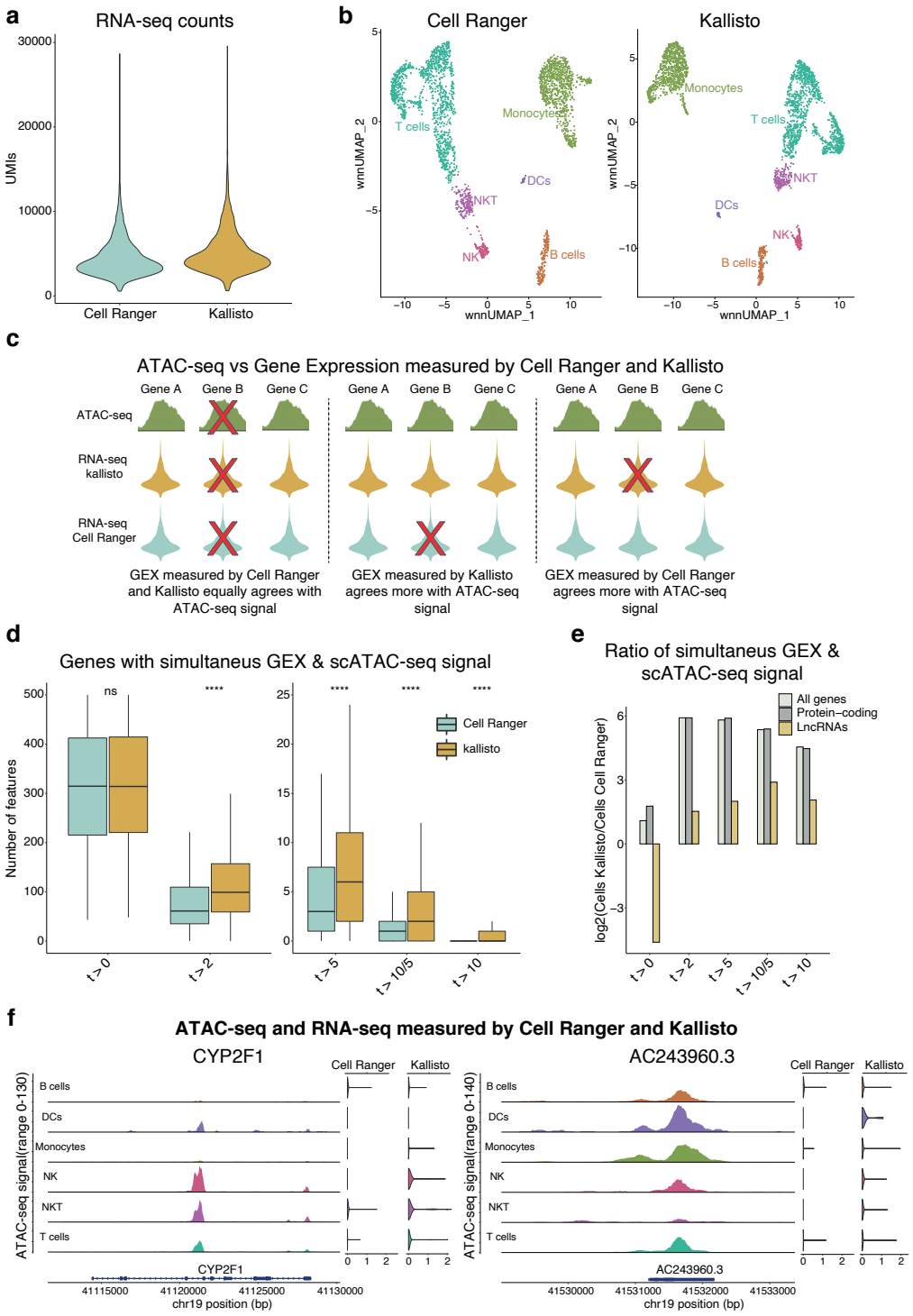

Figure 2. scATAC-seq multiome indicates the optimal preprocessing alternative for lncRNA quantification.

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, 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. **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 (grey) 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).

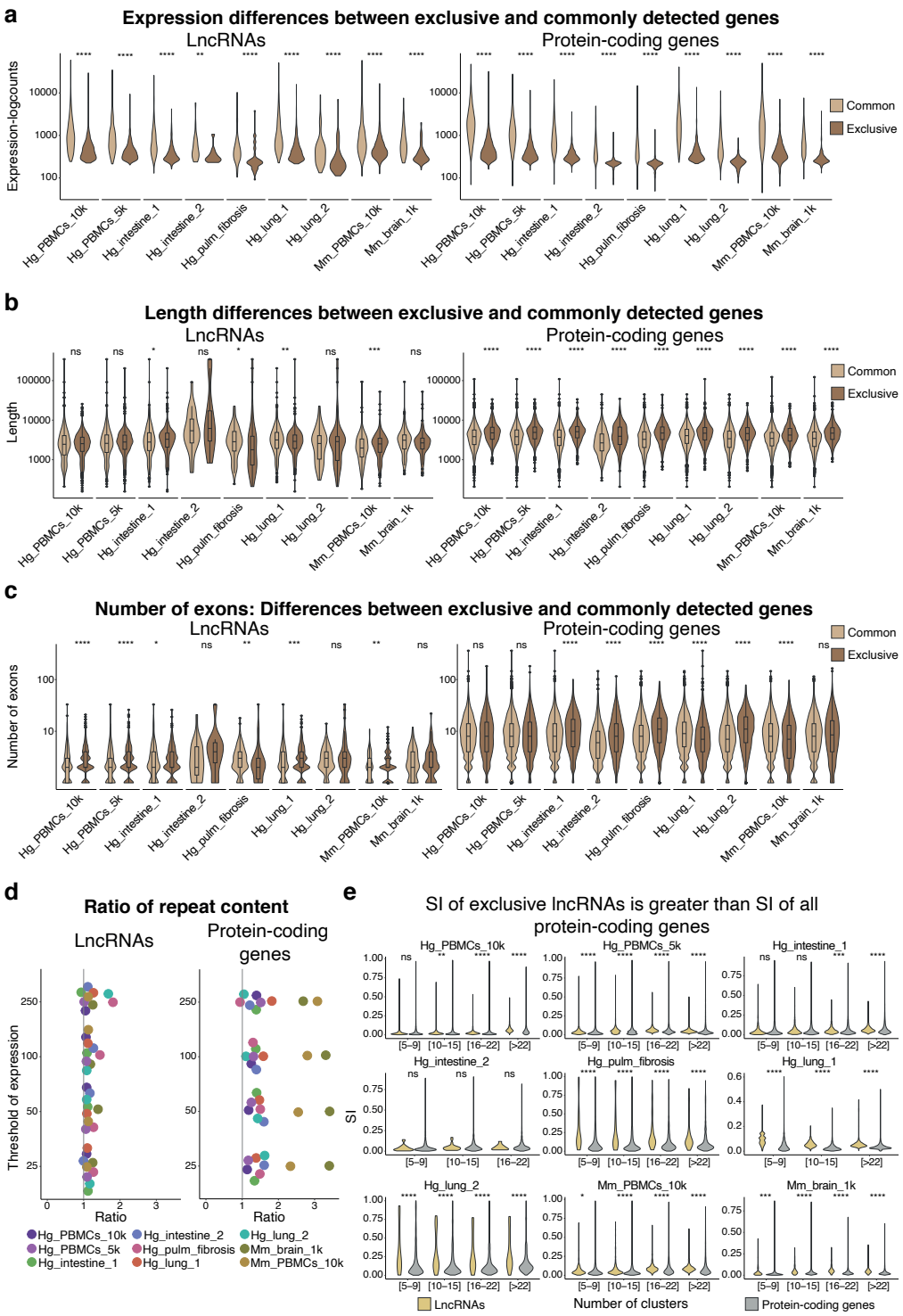

Figure 3. Exclusive and commonly identified lncRNAs share similar characteristics.

a, Normalized expression differences, **b,** Length differences and **c,** Differences in the number of exons of (left) lncRNAs exclusively found by Kallisto compared to lncRNAs commonly identified by all pipelines and (right) protein-coding genes exclusively found by Kallisto compared to protein-coding genes commonly identified by all preprocessing options. In (**a**), (**b**) and (**c**) only genes with more

than 250 counts and present in more than 25 cells were considered. Significance in (**a**), (**b**) and (**c**) was assessed with a two-tailed Wilcoxon test. **d,** Ratio of the sequence covered by repeats of exclusive (left) lncRNAs (right) protein-coding genes divided by the percentage of the sequence covered by repeats of common (left) lncRNAs (right) protein-coding genes. A jitter on the y-axis was included for ease of visualization. Different thresholds on expression for removing lowly-expressed genes were considered, from down to up; more than 25 counts and present in more than 3 cells, more than 50 counts and present in more than 5 cells, more than 100 counts and present in more than 10 cells and more than 250 counts and present in more than 25 cells. **e,** Specificity index (SI) differences to test if the SI of lncRNAs exclusively identified by Kallisto is significantly higher (one-tailed Wilcoxon test) than the SI of protein-coding genes. The SI distributions were calculated across distinct ranges of number and sizes of clusters, from fewer and bigger to more and smaller clusters. Starting with 5-9 clusters, then 10-15 clusters, then 16-22 clusters and finally generating a specially detailed subdivision with at least 23 clusters. Only genes with more than 250 counts and present in more than 25 cells were considered.

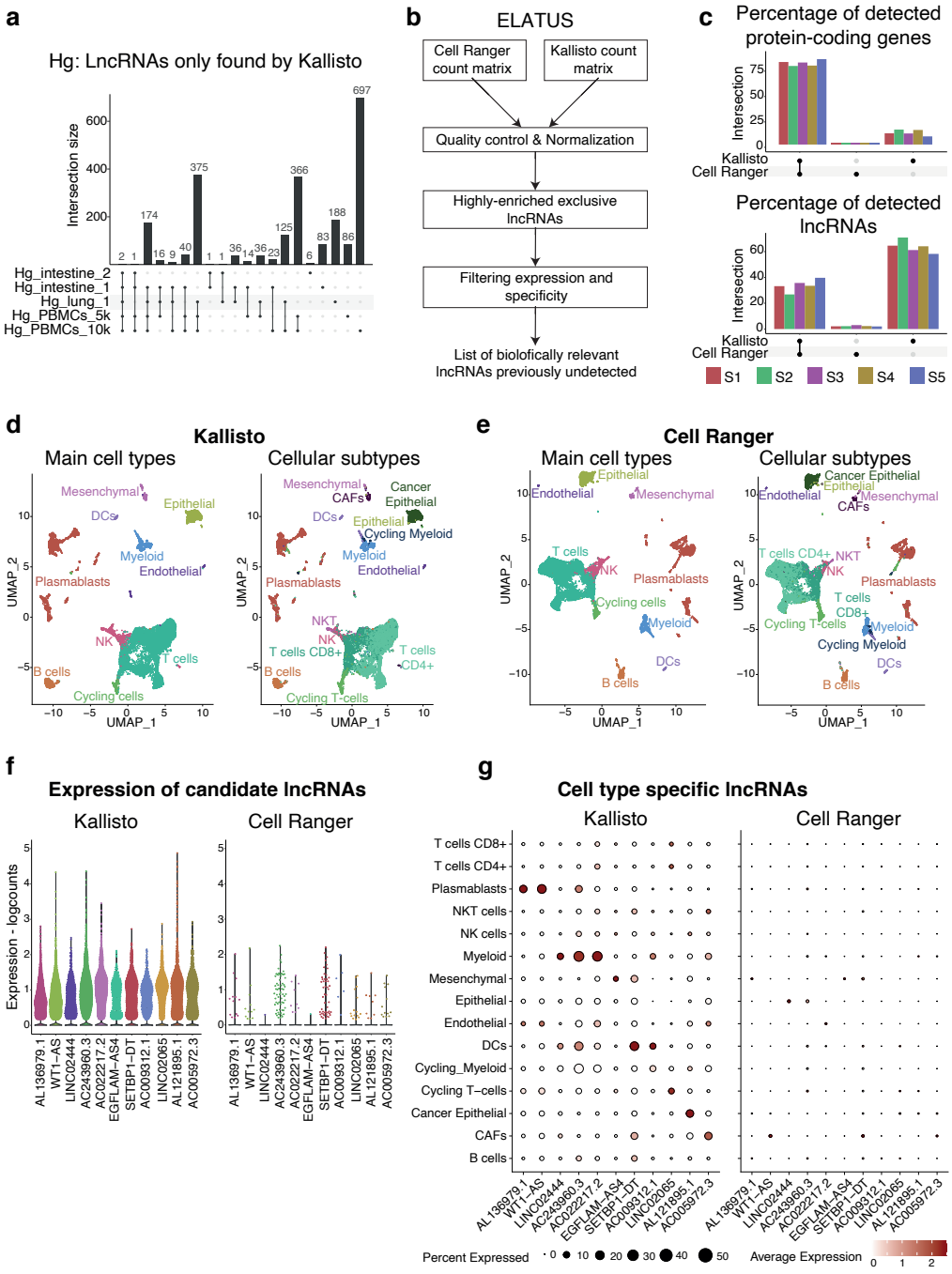

Figure 4. Biologically relevant lncRNAs are uncovered by ELATUS. **a,** UpSet plot displaying the overlap of lncRNAs exclusively found by Kallisto in the human scRNA-seq datasets analyzed. **b,** ELATUS workflow to uncover biologically important lncRNAs previously missed. It started by importing the raw count matrices obtained after preprocessing with both Cell Ranger and Kallisto. Next, there is a quality control step to distinguish empty droplets from cells, filtering of potential multiplets and cells with high mitochondrial content, followed by a normalization and clustering steps. Then, highly expressed lncRNAs exclusively detected by Kallisto were selected and their Specificity Index (SI) was calculated. Finally, 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 were selected as functionally relevant lncRNAs. **c,** UpSet plot displaying, as a percetange, 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.

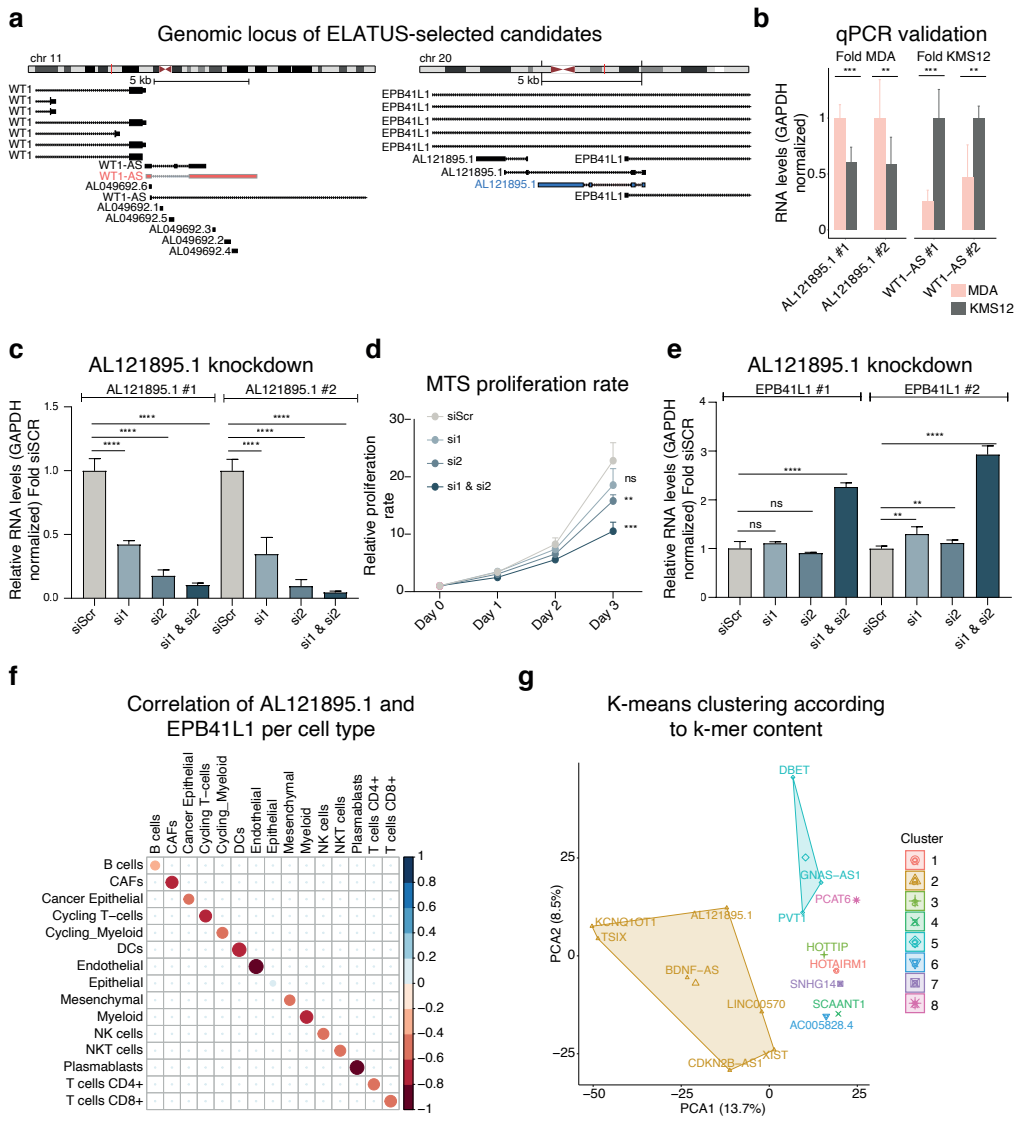

Figure 5. ELATUS-identified *AL121895.1* is a *cis*-repressor that participates in breast cancer proliferation.

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,** qPCR normalized RNA levels 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. **c,** qPCR normalized RNA levels showing the expression of *AL121895.1* on MDA cells after treating them with scramble (siSCR) or performing its knockdown with the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2). **d,** MTS proliferation assay of MDA cells measured during three days treating them with scramble (siSCR) or performing its knockdown with the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2). **e,** qPCR normalized RNA levels showing the expression of *EPB41L1* treating them with scramble (siSCR) or performing its knockdown with

the siRNA1 (si1), siRNA2 (si2) and the combination of both siRNAs (si1 & si2). Two-tailed student t-test for assessing statistical significance in the last four subpanels. **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 previsously describer lncRNAs *cis*-activators and lncRNAs *cis*-repressors.

Supplementary information legends

Supplementary Figure 1. Evolution of the number of protein-coding genes and lncRNAs annotated by GENCODE from 2010 until 2022.

Supplementary Figure 2. Preprocessing choices strongly affect lncRNA detection in a scRNA-seq dataset consisting of 1k brain cells from an E18 mouse.

a, Mapping rate by each pipeline. **b,** Running time per pipeline. **c,** Maximum memory used per pipeline. **d,** Mitochondrial content per cell by each pipeline. **e,** Number of UMI counts per cell by each pipeline. **f,** UpSet plot showing the overlap of retained high-quality cells by each pipeline. **g,** UMAP plots displaying the populations of cell types identified across pipelines. **h,** DotPlots showing across different pipelines, the averaged normalized expression, in each cell type, of canonical marker genes. **i,** Number of detected protein-coding genes per cell across pipelines. **j,** Number of detected lncRNAs per cell across pipelines. **k,** UpSet plot displaying the overlap of highly-expressed protein-coding genes (more than 250 counts and present in more than 25 cells) per pipeline. **i,** UpSet plot displaying the overlap of highly-expressed lncRNAs (more than 250 counts and present in more than 25 cells) per pipeline.

Supplementary Figure 3. Preprocessing choices strongly affect lncRNA detection in a scRNA-seq dataset consisting of 10k human PBMCs from a healthy donor.

a, Running time per pipeline. **b,** Maximum memory used per pipeline. **c,** Mitochondrial content per cell by each pipeline. **d,** DotPlots showing across different pipelines, the averaged normalized expression, in each cell type, of canonical marker genes. **e,** UpSet plot displaying the overlap of highly-expressed protein-coding genes across pipelines in different thresholds. Specifically, in left) more than 25 counts and present in more than 3 cells, middle) more than 50 counts and present in more than 5 cells, right) more than 100 counts and present in more than 10 cells. **f,** UpSet plot displaying the overlap of highly-expressed lncRNAs across pipelines in different thresholds. Specifically, in left) more than 25 counts and present in more than 3 cells, middle) more than 50 counts and present in more than 5 cells, right) more than 100 counts and present in more than 10 cells.

Supplementary Figure 4. Preprocessing choices strongly affect lncRNA detection in an expanded and diverset set of scRNA-seq datasets.

a, Number of UMI counts per cell obtained with Cell Ranger and Kallisto in each dataset. **b,** UpSet plot displaying the percentage of the detected high-cells which are individually or commonly retained by Cell Ranger and Kallisto across datasets. **c,** Number of detected protein-coding genes per cell by Cell Ranger and Kallisto in each dataset. **d,** Number of detected lncRNAs per cell by Cell

Ranger and Kallisto in each dataset. **e,** UpSet plot displaying, for highly-expressed protein-coding genes, the left) percentage of them that have more than 250 counts in more than 25 cells that overlaps between Cell Ranger and Kallisto middle) overlap of them that have more 250 counts in more than 25 cells between Cell Ranger and Kallisto, right) overlap of them that have more 50 counts in more than 5 cells between Cell Ranger and Kallisto. **f,** UpSet plot displaying, for highly-expressed lncRNAs, the left) percentage of them that have more than 250 counts in more than 25 cells that overlaps between Cell Ranger and Kallisto middle) overlap of them that have more 250 counts in more than 25 cells between Cell Ranger and Kallisto, right) overlap of them that have more 50 counts in more than 5 cells between Cell Ranger and Kallisto.

Supplementary Figure 5. scATAC-seq multiome indicates the optimal preprocessing alternative for lncRNA quantification.

a, ATAC-seq counts for each nucleus that fits ATAC-seq quality control thresholds. **b,** Transcription Start Site (TSS) enrichment for each nucleus that fits ATAC-seq quality control thresholds. **c,** Nucleosome signal for each nucleus that fits ATAC-seq quality control thresholds. **d,** Mitochondrial content per nuclei by Cell Ranger and Kallisto. **e,** DotPlots showing across different pipelines, the averaged normalized scRNA-seq expression, in each cell type, of canonical marker genes. **f,** Odds ratio displaying the probability of having more simultaneous activation with Kallisto than with Cell Ranger across different thresholds for defining a gene to be simultaneously activated. Specifically, the x-axis represents the different thresholds used for quantifiying 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. For each nucleus we have considered the expression of all genes (left), only protein-coding genes (middle) and only lncRNAs (right).

Supplementary Figure 6. Exclusive and common lncRNAs have similar expression and length distributions.

a, Normalized expression differences and **b,** Length differences of (left) lncRNAs exclusively found by Kallisto compared to lncRNAs commonly identified by all pipelines and (right) protein-coding genes exclusively found by Kallisto compared protein-coding genes commonly identified by all pipelines. Significance was assessed with a two-tailed Wilcoxon test in (**a**) and (**b**). Thresholds in (**a**) and (**b**) applied for removing poorly-expressed genes were; up) only genes with more than 100 counts and present in more than 10 cells were considered, middle) only genes with more than 50 counts and present in more than 5 cells were considered, down) only genes with more than 25 counts and present in more than 3 cells were considered.

Supplementary Figure 7. Exclusive lncRNAs are highly specific and have a higher number of exons than those commonly detected.
a, Differences in the number of exons of (left) lncRNAs exclusively found by Kallisto compared to lncRNAs commonly identified by all pipelines and (right) protein-coding genes exclusively found by Kallisto compared to protein-coding genes commonly identified by all pipelines. Significance was assessed with a two-tailed Wilcoxon test. Thresholds applied for removing poorly-expressed genes were; up) only genes with more than 100 counts and present in more than 10 cells were considered, middle) only genes with more than 50 counts and present in more than 5 cells were considered, down) only genes with more than 25 counts and present in more than 3 cells were considered. **b,** Percentage of exclusive and common lncRNAs across six SEEKR functional communities [0-5] according to k-mer content of lncRNAs. The percentage of exclusive and common lncRNAs in total also included to evaluate differences (column *ALL*). To remove poorly-expressed genes we considered only genes with more than 250 counts and present in more than 25 cells were considered. **c,** Specificity index (SI) differences to test if the SI of lncRNAs exclusively identified by Kallisto is significantly lower (one-tailed Wilcoxon test) than the SI of protein-coding genes. The SI distributions were calculated across distinct ranges of number and sizes of clusters, from fewer and bigger to more and smaller clusters. Starting with 5-9 clusters, then 10-15 clusters, then 16-22 clusters and finally generating a specially detailed subdivision with at least 23 clusters. To remove poorly-expressed genes we considered only genes with more than 250 counts and present in more than 25 cells were considered.

Supplementary Figure 8. Functional lncRNAs are uncovered by ELATUS.
a, UpSet plot displaying the overlap of lncRNAs exclusively found by Kallisto in the mouse scRNA-seq datasets analyzed. **b,** DotPlots showing the averaged normalized expression, in each cellular subtype, of canonical marker genes when preprocessing with Kallisto and Cell Ranger **c,** UMAPs showing the normalized expression of canonical marker genes when preprocessing with Kallisto **d,** Design of the set of qPCR primers and siRNAs used for *WT1-AS* and *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 **e,** Left) Genomic locus of the protein-coding gene *EPB41L1* and right) Design of the pair of qPCR primers used for the isoform containing most scRNA-seq reads assigned by Kallisto.

Supplementary Figure 9. Exclusive lncRNAs are significantly enriched among CRISPR functionally validated lncRNAs.
Hypergeometric test performed to test the significance of the overlap. Different thresholds on expression for removing lowly-expressed genes were considered,

from left to right; more than 25 counts and present in more than 3 cells (dark blue), more than 50 counts and present in more than 5 cells (blue), more than 100 counts and present in more than 10 (light blue) cells and more than 250 counts and present in more than 25 cells (grey).

Supplementary Figure 10. Conflicts with the intronic annotation when an exon of a gene is overlapped with an intronic region of another gene.
a, ATAC-seq signal and RNA-seq expression, with both Cell Ranger and Kallisto, where part of the mono-exonic *JUNB* is overlapped by an intronic region of *HOOK2*. **b,** ATAC-seq signal and RNA-seq expression, with both Cell Ranger and Kallisto, where the mono-exonic *IGKC* is overlapped by an intronic region of *AC244205.1*.

Supplementary Table 1.
Canonical markers used for distinguishing main cell types.

Supplementary Table 2.
List of lncRNAs exclusively detected by Kallisto identified by public CRISPRi screenings.

Supplementary Table 3.
List of 260 lncRNAs that are exclusively detected with Kallisto and exhibit characteristics of functional lncRNAs.

Supplementary Table 4.
List of 10x Genomics scRNA-seq datasets used.