Long-read sequencing transcriptome quantification with lr-kallisto

Rebekah K. Loving[1], Delaney K. Sullivan[1,2], A. Sina Booeshagi[3], Fairlie Reese[4,5], Elisabeth Rebboah[4,5], Jasmine Sakr[4,5], Narges Rezaie[4,5], Heidi Y. Liang[4,5], Ghassan Filimban[4,5], Shimako Kawauchi[4], Conrad Oakes[1], Diane Trout[1], Brian A. Williams[1], Grant MacGregor[4], Barbara J. Wold[1], Ali Mortazavi[4,5], and Lior Pachter[1,6]

[1]Division of Biology and Biological Engineering, California Institute of Technology, USA
[2]UCLA-Caltech Medical Scientist Training Program, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA
[3]Department of Bioengineering, University of California, Berkeley, Berkeley, USA
[4]Developmental and Cell Biology, University of California Irvine, Irvine, USA
[5]Center for Complex Biological Systems, University of California Irvine, Irvine, USA
[6]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, USA

RNA abundance quantification has become routine and affordable thanks to high-throughput "short-read" technologies that provide accurate molecule counts at the gene level. Similarly accurate and affordable quantification of definitive full-length, transcript isoforms has remained a stubborn challenge, despite its obvious biological significance across a wide range of problems. "Long-read" sequencing platforms now produce data-types that can, in principle, drive routine definitive isoform quantification. However some particulars of contemporary long-read datatypes, together with isoform complexity and genetic variation, present bioinformatic challenges. We show here, using ONT data, that fast and accurate quantification of long-read data is possible and that it is improved by exome capture. To perform quantifications we developed lr-kallisto, which adapts the kallisto bulk and single-cell RNA-seq quantification methods for long-read technologies.

Correspondence: ***Barbara Wold (woldb@caltech.edu), Ali Mortazavi (ali.mortazavi@uci.edu) and Lior Pachter (lpachter@caltech.edu)***

Introduction

Advances in long-read RNA sequencing are facilitating transcript discovery, annotation improvements, and detection of isoform switching, thanks to reductions in cost and decreasing error rates as the technologies mature (1–3). Specifically, long-read RNA-seq can readily detect gene fusion transcripts and other expressed rearrangements in cancer (4), and isoform switching of biological consequence across development (5, 6). In translational genomics, precision medicine workflows are increasingly including gene and transcript ontology. These capabilities depend, in part, on accurate annotation of the genomes and transcriptomes of human and model organisms, though they remain incomplete (7, 8). Improvements in long-read sequencing now allow for much needed refinement of annotations for human and model organisms, coupled with rapid generation of genomes and annotations for non-model organisms (9). Importantly, while annotation is mainly facilitated by transcript discovery, quantification of isoforms is critical for filtering and thresholding steps that are prerequisites for resolving locus structure and quantifying their expression products (10).

While recent increases in affordability and sequence quality are bringing full-isoform quantification within reach, the long-read platforms are still rapidly changing and less mature than short-read technologies (2). For example, Oxford Nanopore Technology (ONT) sequencing has evolved over many versions of chemistry in the library preparation kits, pores, and signal processing algorithms. This has resulted in a range of ONT data with various error profiles and error distributions within the sequences. Of the quantification tools that have been developed so far (11–19), many are optimized for performance with a given generation of long-read data and are now antiquated, in both accuracy and efficiency, for processing the low error rate ONT data currently being produced. Moreover, many methods are based on the assumption of near uniform distribution of sequencing error along reads; we found, as have others (20), that this does not hold in practice. Furthermore, some ONT sequencing biases have now been described, including non-uniformly distributed sequencing error and sequence influenced error, such as higher GC content and repeat regions increasing sequencing/base calling error (21).

By contrast, several accurate and efficient tools have been developed for short read RNA-seq preprocessing (22–27). However, even with the recent significant reduction in the long-read RNA-seq error rates to $\sim$0.5%, sequencing errors remain informatically problematic and are comparatively much higher than the $\sim$0.01% in short-read RNA-seq. This makes the application of the fastest pseudoalignment methods (25, 27) to long-reads nontrivial. Our approach, which builds on kallisto (23–25, 28) and which we term lr-kallisto, demonstrates the feasibility of pseudoalignment for long-reads; we show via a series of results on both biological and simulated data that lr-kallisto retains the efficiency of kallisto thanks to pseudoalignment, and is accurate on long-read data. Furthermore, we show that lr-kallisto is robust to error rates, making it suitable also for the analysis of previously published older long-read sequencing data.

Results

To assess the accuracy of lr-kallisto with respect to data from the current Oxford Nanopore Technologies platform (see Materials and Methods) we generated a deep coverage, high fidelity dataset using long-read sequence and an Illumina

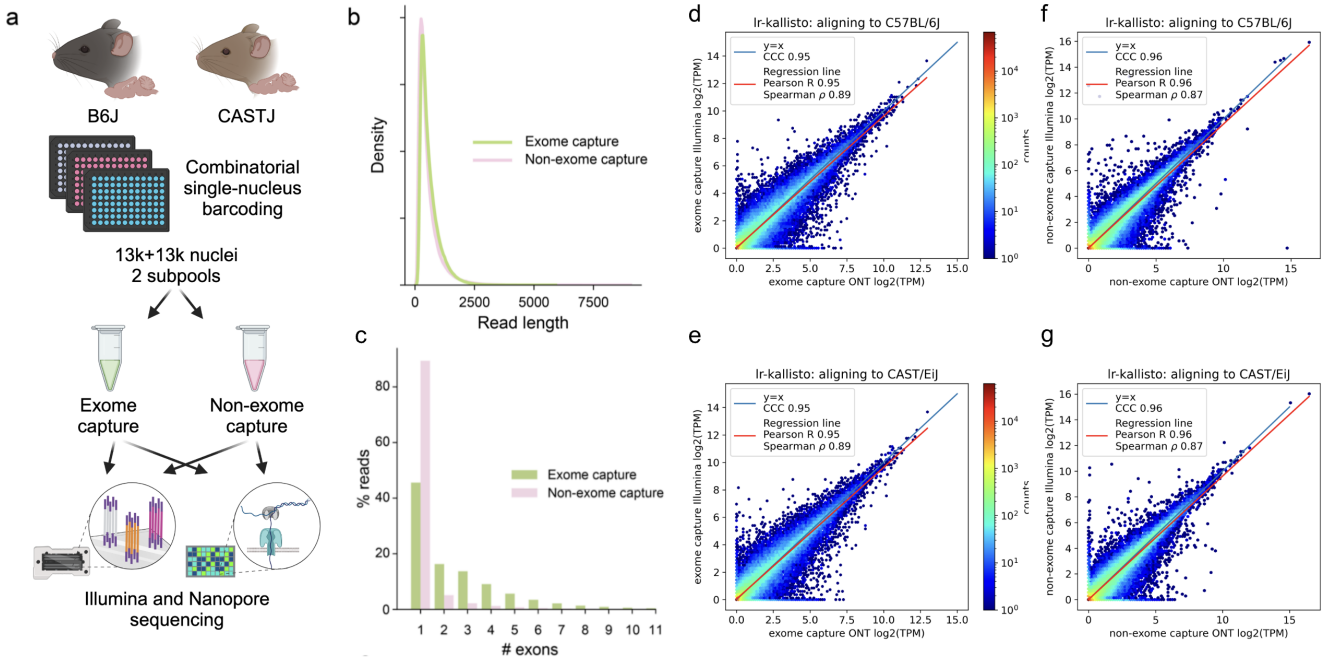

Fig. 1. lr-kallisto demonstrates high concordance between Illumina and ONT. (a) Experimental overview for comparison of exome capture vs. non-exome capture LR-Split-seq libraries. (b) Kernel density estimations for read length distributions by capture strategy. (c) Percentage of demultiplexed reads by number of exons in each read between exome and non-exome capture. (d) lr-kallisto pseudobulk quantifications of exome capture for the C57BL/6J sample. (e) lr-kallisto pseudobulk quantifications of exome capture for the CAST/Eij sample. (f) lr-kallisto pseudobulk quantifications of non-exome capture for the C57BL/6J sample. (g) lr-kallisto pseudobulk quantifications of non-exome capture for the CAST/Eij sample. Concordance Correlation Coefficient (CCC), Pearson, and Spearman correlations are shown for each comparison.

short-read sequence SPLiT-Seq nuclei of the left cortices of two mouse strains as part of the IGVF consortium (29). Specifically, 4 biological replicates (2 males and 2 females) were assayed from both C57BL6/J and CAST/EiJ mice, all at 10 weeks of age, with libraries generated with and without targeted exome capture of all mouse protein coding exons using the Twist Biosciences Mouse exome panel of 215,000 probes (Fig. 1a; see Methods). Thus our exome capture transcriptome will be enriched for reads overlapping one or more coding exons in the same cell. This platform and experimental design was chosen to produce starting data with a highly relevant sequencing error profile for two very well characterized genomes whose natural genetic variation between strains is similar to that found within individual human genomes. This also sets the stage for using lr-kallisto to study natural genetic variation.

We found no effective difference in read lengths with reads generated from exome capture as opposed to non-exome capture libraries (Fig. 1b), though the exome capture library showed a smaller fraction of mono-exonic reads (Fig. 1c). This indicated that exome capture is an effective approach to increasing the transcriptome complexity of libraries. The Illumina and ONT sequenced libraries displayed high transcript abundance concordance after quantification with lr-kallisto (Fig. 1d-g), showing that lr-kallisto accurately quantifies transcripts from long-reads, as well as demonstrating that deeply sequenced ONT libraries are suitable for high accuracy quantification. The concordance correlation coefficients (CCC), which measure how close the ONT and Illumina quantifications are to being identical, were high for both the exome capture and non-exome capture libraries (0.95 and 0.96, respectively). Importantly, when comparing all long-reads that were subject to exome capture versus those that were not, we observed a 3-fold increase in the percentage of spliced reads aligning (Supplementary Fig. 1a). Thus, we find that exome capture reads help overcome the limitations of RNA sampling in the nucleus, including expected reads from unspliced precursor transcripts. The effect, as others have noted (30) is to provide more informative reads to study full-length, spliced transcript isoform usage at lower cost. Furthermore, lr-kallisto outperforms Bambu (15), IsoQuant (17), and Oarfish (16) with respect to CCC, Pearson correlation, and Spearman correlation (Supplementary Fig. 1b). In particular, the lr-kallisto CCC is 0.95 versus 0.82 for the recently published Oarfish long-read quantification tool. We found that lr-kallisto also outperforms Bambu (CCC = 0.86) and IsoQuant (CCC = 0.78) (Supplementary Fig. 1b), which have previously been shown to outperform other long-read quantification methods (2, 31). In addition to being more accurate than other methods, lr-kallisto is also more computationally efficient (Supplementary Fig. 1c). Note that the dramatic difference in time scales between PacBio and ONT is due to the number of reads in the ONT datasets being much higher, in general.

Importantly, lr-kallisto can be used for both high-throughput bulk RNA-seq as well as single-cell or single-nuclei RNA-seq datasets (Supplementary Fig. 1d), and is not only faster than other tools, but also benefits from the low-memory requirements of kallisto (23, 28). For single-nuclei RNA-seq processing, we used splitcode (32) to extract nuclei barcodes, umis, and reads from the raw ONT reads and then pseudoaligned and quantified the reads with lr-kallisto (see

Methods). 100% of barcodes from the ONT reads that passed filtering were also found in Illumina sequenced reads (Supplementary Fig. 1d). Increased UMI depth per nucleus yields higher Spearman correlations, indicating that with deeper sequencing depth, short and long read correlations will only improve (Supplementary Fig. 1dI). To assess the observed correlations between Nanopore and Illumina, we evaluated random oligo vs 3' priming in Illumina sequenced reads and ONT sequenced reads separately, in the same fashion, finding lower correlations (majority of nuclei having a Spearman ρ between 0.10 and 0.30) than between ONT sequenced reads and Illumina sequenced reads (Supplemental Fig. 1dII-III).

We also examined the concordance between the exome capture and non-exome capture in both long and short reads, and found it to be only CCC = 0.88, highlighting the distortion resulting from the coupling of exome capture with a mix of 3'-end and randomly primed read sequencing that is characteristic of Parse (Supplementary Fig. 1e).

The lr-kallisto quantification results are corroborated when comparing its performance to other tools on previously published data that is less deeply sequenced. In a comparison of Illumina and ONT on the HCT116 cancer cell line dataset generated by SG-NEx (33), we found that lr-kallisto could accurately quantify isoform level expression, in performance comparisons constituting two replicates of direct cDNA and direct RNA (Supplementary Fig. 2a). The CCC performance of lr-kallisto exceeded that of Oarfish, evaluated on this data in (16). Spearman correlations were lower overall in this dataset, indicating poor data quality, perhaps due to the lower coverage and higher sequencing error rate. We also compared lr-kallisto's performance on direct RNA to direct cDNA (Supplementary Fig. 2b). We also found better performance with direct cDNA versus direct dRNA, and hypothesize that this is likely due to $\sim$4 times the depth of coverage for replicate 4 in direct cDNA (7,656,893 reads) vs the direct RNA replicate 4 (1,896,643 reads), whereas replicate 3 direct cDNA (873,077 reads) vs direct RNA (1,185,183 reads) does not have the increased depth of coverage. We also compared lr-kallisto to Bambu, IsoQuant, and Oarfish on a previously sequenced mouse cortex PacBio dataset (Supplementary Fig. 2c). On this dataset (34, 35), which has an error rate of 12.4% (see Methods) and a different error profile with errors more uniformly distributed along transcripts, we found similar performance between programs with lr-kallisto slightly outperforming other tools in CCC.

We benchmarked lr-kallisto's stability and robustness compared to other long-read quantification tools across species, platforms, and protocols, by evaluating lr-kallisto's performance, along with Bambu, IsoQuant, and Oarfish using the LRGASP's challenge 2 benchmark (2) of long-read quantification tools (Fig. 2). For our benchmarking, we chose to focus on the Mouse ES data, as it had lower sequencing error rates across 3 out of the 4 protocol/platform combinations, thereby serving as the closer proxy for current long-read data. We found that Bambu, IsoQuant, lr-kallisto, and Oarfish all achieved reasonably high CCCs between replicates, both with respect to abundance estimates (Fig. 2a), and variability between isoforms (Fig. 2b). For completeness, we also compared the performance of lr-kallisto to Bambu, IsoQuant, Oarfish using the metrics of the LRGASP paper (Supplementary Fig. 3). Resolution Entropy (RE) is a measure of how well a tool uniformly quantifies at different expression levels. Irreproducibility Measure (IM) is a measure of how reproducibly the tool quantifies expression across replicates, i.e., whether the coefficient of variation between replicates is low. Consistency Measure (CM) is a measure of how consistent the tool is at detecting expressed transcripts, assuming that transcripts should be expressed simultaneously across replicates, and ACVC is the Area under the Coefficient of Variation Curve, which again assumes that for a given mean expression level across replicates the coefficient of variation should be low. We found that lr-kallisto performs as well as other programs on these stability and robustness measures. The variability that we found in quantifications of replicates can be explained by variable depth of sequencing between the replicates and between the protocols and platforms (2). The notable difference in ONT cDNA is due in part to a sequencing error rate of $\sim$12%, which is characteristic of data obtained in earlier ONT platform versions (36).

We assessed the performance of lr-kallisto using simulated data across a range of sequencing error profiles, and compared with results on the same simulated data for five other widely used or recently published programs. We used simulations generated by (17) who used the IsoSeqSim simulator (see Data and Code Availability) to generate PacBio reads (6 million *Mus musculus* reads with $\sim$1.4% sequencing error rate), and NanoSim (18) to generate ONT.R10.4 reads (30 million *Mus musculus* reads with $\sim$2.8% sequencing error rate). The PacBio IsoSeqSim Simulation (Fig. 3a) demonstrates lr-kallisto's high accuracy compared to the currently leading benchmarked long-read quantification tools Bambu, IsoQuant, and Oarfish, with lr-kallisto achieving a CCC of 0.98, vs 0.90, 0.91, and 0.99, respectively (2, 31). Furthermore, in the ONT NanoSim R10.4 Simulation (Fig. 3b), lr-kallisto ties for the highest CCC of 0.97, vs 0.88 and 0.91, respectively.

We performed additional comparative evaluations of Bambu, IsoQuant, lr-kallisto, and Oarfish on a more extensive set of simulations to understand the strengths and weaknesses these tools when confronted with different sequencing error challenges (Supplementary Fig. 4). We found that lr-kallisto and IsoQuant were both robust to indel and substitution profiles simulated to match PacBio sequencing data and uniformly distributed. IsoQuant was also robust to uniformly distributed sequencing errors with indel and substitution profiles matched to ONT, whereas lr-kallisto performance degraded at higher ONT error rates in this simulation (Supplementary Fig. 4a). In particular, this highlights lr-kallisto's sensitivity to the unrealistic combination of uniform sequencing error distribution and higher rate of insertion errors in ONT versus PacBio.

In another ONT simulation generated with NanoSim to produce reads with an 11.2% error rate (see Data and Code Availability), lr-kallisto achieved a CCC of 0.31 on

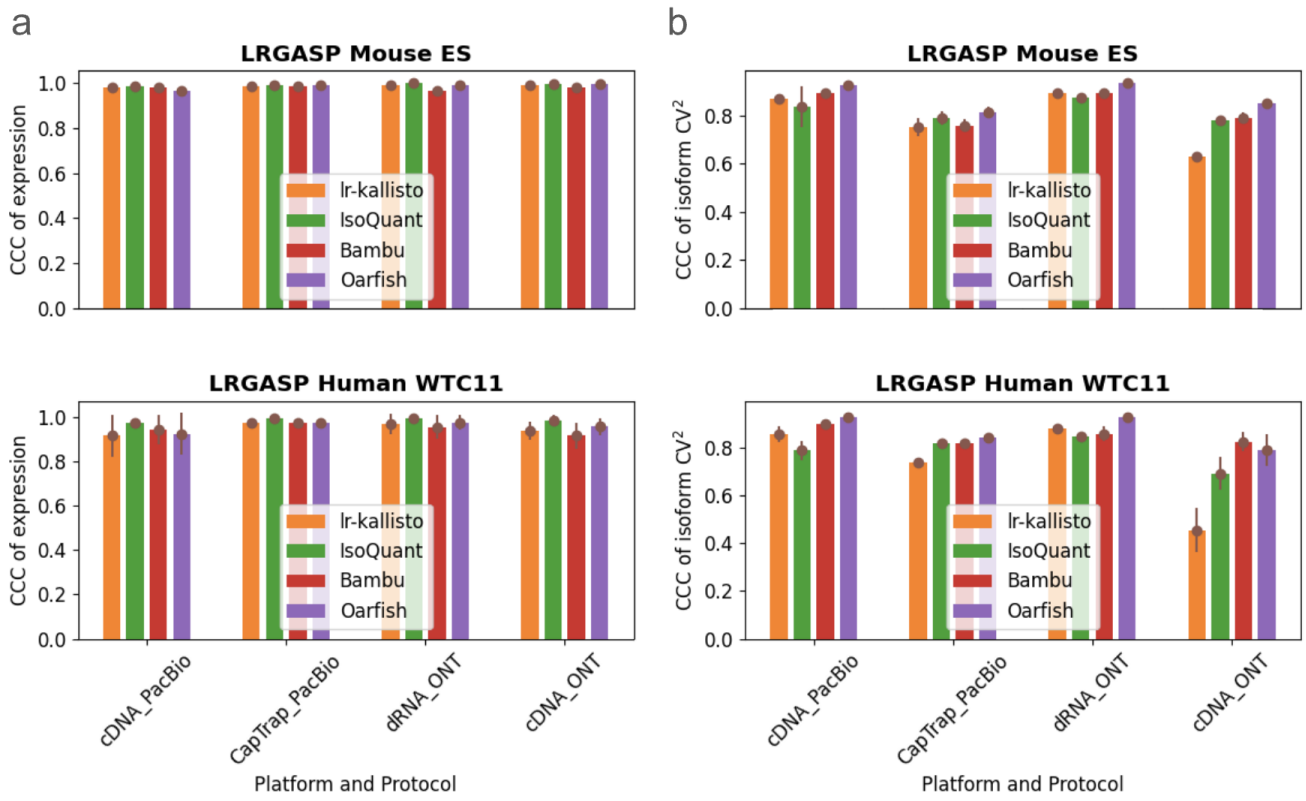

Fig. 2. Comparison of Bambu, IsoQuant, lr-kallisto, and Oarfish in (a) abundance estimates as measured by CCC of expression and (b) variability between isoforms as measured by CCC of isoform CV^2, with 90% CI to measure consistency and reproducibility among replicates between the tools.

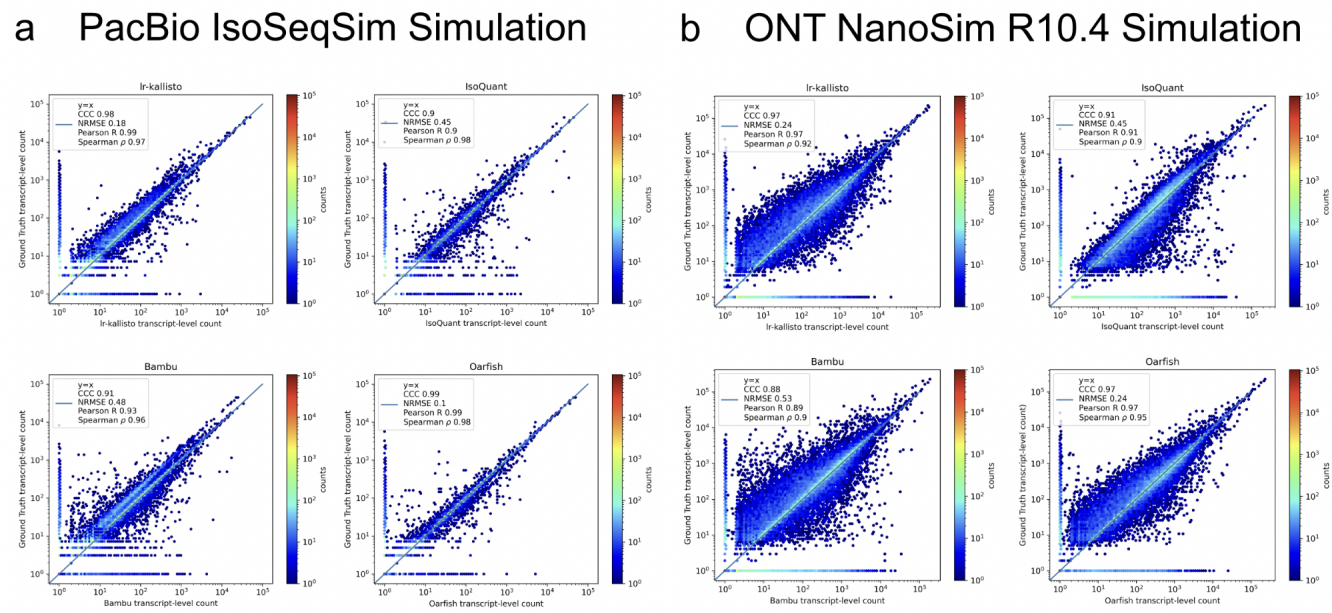

Fig. 3. lr-kallisto is highly accurate in simulations with error up to ∼3%. A comparison of performance of Bambu, IsoQuant, lr-kallisto, and Oarfish on PacBio (left-hand side) and ONT (right-hand side) simulations with Concordance Correlation Coefficient (CCC), Normalized Root Mean Squared Error, and Pearson's and Spearman's correlation coefficients reported.

all transcripts, outperforming IsoQuant (CCC = 0.28), and underperforming Bambu (CCC = 0.51), and Oarfish (CCC = 0.55) (Supplementary Fig. 4b). This was also the case at a higher error rate (15.2%), with lr-kallisto continuing to outperform IsoQuant and underperform Bambu and Oarfish (Bambu CCC = 0.53, IsoQuant CCC = 0.32, lr-kallisto CCC = 0.34, Oarfish CCC = 0.58) (Supplementary Fig. 4c).

The performance of lr-kallisto benefits from quantification with respect to a de Bruijn graph (28). We tested whether and to what extent changing the *k*-mer length default in lr-kallisto to 63 bp long vs 31 bp long in the reference transcriptome de Bruijn graph creates a less connected and less

complex structure (Supplementary Fig. 5). In this example, of the Pax2 gene, we find that a change of *k*-mer length simplifies the T-DBG with the reduction of a single node and 2 edges. However, when we scale this out to just the first 1000 transcripts listed in the LRGASP basic gencode human annotation, we found a reduction from 3,698 nodes using the 31 *k*-mer T-DBG to 2,708 nodes using the 63 *k*-mer T-DBG and 4,687 edges to 3,238 edges, respectively. Furthermore, the largest connected T-DBG graph component in the 63 *k*-mer T-DBG is composed of 12.59% of the bp vs 65.90% in the 31 *k*-mer T-DBG. We believe that the selection of higher quality, low sequencing error regions from the reads by the 63 *k*-mer T-DBG, combined increasing the probability of uniquely mapping, or at the very least mapping to a transcript compatibility class with less transcripts, is producing more accurate and more efficient pseudoalignment.

Discussion

With Oxford Nanopore sequencing becoming more accessible due to low entry costs and reduced sequencing error rate (37), long-read sequencing is advancing our ability to decipher the complexity of transcriptomes. Increasing throughput now makes it possible to not only perform discovery with long-read sequencing, but also to accurately quantify transcript abundances, and we have shown that results comparable to short-read sequencing can be achieved at reasonable cost with exome capture, and with high accuracy quantification using lr-kallisto. Exome capture will be especially helpful for filtering out intronic reads that would be otherwise sequenced in (single-)nucleus data, as nuclei are replete with intron lariats and partially processed transcripts. lr-kallisto is highly accurate in producing quantification results on data with less than 10% sequencing error rate comparable to those with short-read sequencing. This makes lr-kallisto immediately useful for current long-read sequencing transcriptome projects, although performance will not be as good on legacy higher error long-read sequencing datasets.

Furthermore, as described in Methods, lr-kallisto is useful for long-read sequencing of single-cell and single-nucleus RNA-seq libraries when coupled with tools designed for barcode discovery (32, 38). Furthermore, lr-kallisto is comptible with translated pseudoalignment, which can be useful for detection of viruses (39).

Finally, in this work we have focused on quantification. However, lr-kallisto can also be used, in principle, for transcript discovery. In particular, reads that do not pseudoalign with lr-kallisto can be assembled to construct contigs from unannotated, or incompletely annotated, transcripts.

Data and code availability

The LRGASP data can be accessed from the accessions and ftp links listed in the data folder of https://github.com/pachterlab/LSRRSRLFKOTWMWMP_2024.
IGVF Bridge exome capture and non-exome capture can be accessed from the IGVF portal with the accession IDs in the provided table.

Accession ID	Subpool Name	Read Type
IGVFDS4803WKTQ	B01_13G	Nanopore
IGVFDS9445YYVB	B01_13H	Nanopore
IGVFDS9522BMQK	B01_13G	Illumina
IGVFDS0356VCIO	B01_13H	Illumina

Table 1. IGVF Bridge exome capture and non-exome capture accession IDs.

The HCT116 cell line SG-NEx data was accessed on March 13, 2024 at https://registry.opendata.aws/sg-nex-data. The lr-kallisto method is available via release 0.51 of kallisto at https://github.com/pachterlab/kallisto.

We used bambu v3.4.1, IsoQuant v3.3.0, and oarfish v0.5.1 (with the exception of analysis of HCT116 data). In the initial version of the preprint, oarfish (v0.3.1 and v0.4.0) were used and the simulation data was run with samtools sort (genome coordinate sorting), causing overcounting in oarfish's performance due to oarfish's use of consecutive alignments of the same read filtering; this has been updated in this version of the manuscript. Simulation data is available at https://zenodo.org/records/11201284. Processed abundance matrices for Figures 1-3 are available at https://zenodo.org/records/13755772. Code for reproducing the results and figures in the manuscript is available at https://github.com/pachterlab/LSRRSRLFKOTWMWMP_2024.

Author contributions

The lr-kallisto project was conceived by RKL and LP and the lr-kallisto method was developed and implemented by RKL. Benchmarking was conducted by RKL. The exome capture / non-capture experiment was conceived by AM, BWo and BWi. The experiment, data generation and curation was supervised by AM. Experiments were conducted by ER, HL, GF, SK and GM. Data curation was performed by FR, JS, DT and NR. Analysis of the data was conducted by RKL, FR and LP. RKL, ASB, and DKS developed the lr-kallisto single-cell workflow including updates to seqspec and splitcode. Supplementary data analysis was performed by RKL, LP and CO. Software testing and release was performed by RKL and DKS. The manuscript was drafted by RKL and LP. LP, RKL, AM, BW, FR, DKS and CO commented on and edited the manuscript. All authors approved the manuscript. LP supervised the lr-kallisto project with BW.

Acknowledgments

This work was partially supported by UM1 HG012077 to A.M., B.J.W., and L.P. as well as a United States Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0020347 to R.K.L. D.K.S. was supported by the UCLA-Caltech Medical Scientist Training Program (NIH NIGMS training grant T32 GM008042). We thank Zahra Zare Jousheghani, Noor Pratap Singh, and Rob Patro for

Accession ID	File Name
IGVFDS4705QPIK	b01_nanopore_13G_single_cell_k63_both_mm39
IGVFDS7467TPQO	b01_nanopore_13G_single_cell_k63_polyT_mm39
IGVFDS3821ZEWS	b01_nanopore_13G_single_cell_k63_randO_mm39
IGVFDS1377KBXL	b01_next1_13G_single_cell_k31_both_mm39
IGVFDS2498XYWS	b01_next1_13G_single_cell_k31_polyT_mm39
IGVFDS9180SYAE	b01_next1_13G_single_cell_k31_randO_mm39
IGVFDS4019MYIG	b01_nanopore_13G_bulk_k63_casteij
IGVFDS6540HMFT	b01_nanopore_13G_bulk_k63_mm39
IGVFDS3833XYEY	b01_nanopore_13H_bulk_k63_casteij
IGVFDS5673HQEN	b01_nanopore_13H_bulk_k63_mm39
IGVFDS2760LQIX	b01_next1_13G_bulk_k31_casteij
IGVFDS9744VNMR	b01_next1_13G_bulk_k31_mm39
IGVFDS0231GDWH	b01_next1_13H_bulk_k31_casteij
IGVFDS1622ABWA	b01_next1_13H_bulk_k31_mm39

Table 2. IGVF Bridge exome capture and non-exome capture processed accession IDs.

comments on consistency and version control following the first version of this manuscript on bioRxiv.

Bibliography

1. Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1):30, 2020.
2. Francisco J Pardo-Palacios, Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E Loveland, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv*, July 2023. doi: 10.1101/2023.07.25.550582.
3. Fairlie Reese, Brian Williams, Gabriela Balderrama-Gutierrez, Dana Wyman, Muhammed Hasan Çelik, Elisabeth Rebboah, Narges Rezaie, et al. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. *bioRxiv*, May 2023. doi: 10.1101/2023.05.15.540865.
4. Yoshitaka Sakamoto, Sarun Sereewattanawoot, and Ayako Suzuki. A new era of long-read sequencing for cancer genomics. *Journal of Human Genetics*, 65(1):3–10, 2020.
5. Chaoyang Wang, Zhuoxing Shi, Qingpei Huang, Rong Liu, Dan Su, Lei Chang, Chuanle Xiao, and Xiaoying Fan. Single-cell analysis of isoform switching and transposable element expression during preimplantation embryonic development. *PLoS Biology*, 22(2):e3002505, 2024.
6. Livius Penter, Mehdi Borji, Adi Nagler, Haoxiang Lyu, Wesley S Lu, Nicoletta Cieri, Katie Maurer, et al. Integrative genotyping of cancer and immune phenotypes by long-read sequencing. *Nature Communications*, 15(1):32, 2024.
7. David Zhang, Sebastian Guelfi, Sonia Garcia-Ruiz, Beatrice Costa, Regina H Reynolds, Karishma D'Sa, Wenfei Liu, et al. Incomplete annotation has a disproportionate impact on our understanding of mendelian and complex neurogenetic disorders. *Science Advances*, 6(24), 2020. doi: 10.1126/sciadv.aay8299.
8. Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, et al. GENCODE 2021. *Nucleic Acids Research*, 49 (D1):D916–D923, 2021.
9. Peter E Warburton and Robert P Sebra. Long-read DNA sequencing: recent advances and remaining challenges. *Annual Review of Genomics and Human Genetics*, 24:109–132, 2023.
10. David E Cook, Jose Espejo Valle-Inclan, Alice Pajoro, Hanna Rovenich, Bart PHJ Thomma, and Luigi Faino. Long-read annotation: Automated eukaryotic genome annotation based on long-read cdna sequencing. *Plant Physiology*, 179(1):38–54, 2019.
11. Alison D Tang, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks. Full-length transcript characterization of sf3b1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature communications*, 11(1):1438, 2020.
12. Luyi Tian, Jafar S Jabbari, Rachel Thijssen, Quentin Gouil, Shanika L Amarasinghe, Oliver Voogd, Hasaru Kariyawasam, Mei RM Du, Jakob Schuster, Changqing Wang, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome biology*, 22:1–24, 2021.
13. Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Shan Jiang, Sorena Rahmanian, Stefania Forner, Dina Matheos, Weihua Zeng, Brian Williams, Diane Trout, et al. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. *Biorxiv*, page 672931, 2019.
14. Matthias Lienhard, Twan van den Beucken, Bernd Timmermann, Myriam Hochradel, Stefan Boerno, Florian Caiment, Martin Vingron, and Ralf Herwig. Isotools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics*, 39(6):btad364, 2023.
15. Ying Chen, Andre Sim, Yuk Kei Wan, Keith Yeo, Joseph Jing Xian Lee, Min Hao Ling, Michael I Love, and Jonathan Göke. Context-aware transcript quantification from long-read RNA-seq data with bambu. *Nature Methods*, 20(8):1187–1195, 2023.
16. Zahra Zare Jousheghani and Rob Patro. Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. *bioRxiv*, March 2024. doi: 10.1101/2024.02.28.582591.
17. Andrey D Prjibelski, Alla Mikheenko, Anoushka Joglekar, Alexander Smetanin, Julien Jarroux, Alla L Lapidus, and Hagen U Tilgner. Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*, 41(7):915–918, 2023.
18. Chen Yang, Justin Chu, René L Warren, and Inanç Birol. Nanosim: nanopore sequence

read simulator based on statistical characterization. *GigaScience*, 6(4):gix010, 2017.

19. Michal Kabza, Alexander Ritter, Ashley Byrne, Kostianna Sereti, Daniel Le, William Stephenson, and Timothy Sterne-Weiler. Accurate long-read transcript discovery and quantification at single-cell resolution with isosceles. *bioRxiv*, pages 2023–11, 2023.
20. Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18): 3094–3100, 2018.
21. Clara Delahaye and Jacques Nicolas. Sequencing DNA with nanopores: Troubles and biases. *PloS One*, 16(10):e0257521, 2021.
22. Benjamin Kaminow, Dinar Yunusov, and Alexander Dobin. STARsolo: Accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv*, 2021. doi: 10.1101/2021.05.05.442755.
23. Delaney K Sullivan, Kyung Hoi Joseph Min, Kristján Eldjárn Hjörleifsson, Laura Luebbert, Guillaume Holley, Lambda Moses, Johan Gustafsson, et al. Kallisto, bustools, and kb-python for quantifying bulk, single-cell, and single-nucleus RNA-seq. *bioRxiv*, November 2023. doi: 10.1101/2023.11.21.568164.
24. Páll Melsted, A Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, Kyung Hoi Joseph Min, Eduardo da Veiga Beltrame, et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, 39(7):813–818, 2021.
25. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016.
26. Hirak Sarkar, Avi Srivastava, Mohsen Zakeri, Scott K Van Buren, Naim U Rashid, Michael I Love, and Rob Patro. Accurate, efficient, and uncertainty-aware expression quantification of single-cell RNA-seq data. *bioRxiv*, 2020. doi: 10.6084/m9.figshare.13198100.
27. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4): 417–419, 2017.
28. Kristján Eldjárn Hjörleifsson, Delaney K Sullivan, Guillaume Holley, Páll Melsted, and Lior Pachter. Accurate quantification of single-nucleus and single-cell RNA-seq transcripts. *bioRxiv*, 2022. doi: 10.1101/2022.12.02.518832.
29. IGVF Consortium. The impact of genomic variation on function (IGVF) consortium. *ArXiv*, 2023. doi: 10.1101/2023.03.28.533945.
30. Tyler Landrith, Bing Li, Ashley A Cass, Blair R Conner, Holly LaDuca, Danielle B McKenna, Kara N Maxwell, Susan Domchek, Nichole A Morman, Christopher Heinlen, et al. Splicing profile by capture rna-seq identifies pathogenic germline variants in tumor suppressor genes. *NPJ precision oncology*, 4(1):4, 2020.
31. Xueyi Dong, Mei RM Du, Quentin Gouil, Luyi Tian, Jafar S Jabbari, Rory Bowden, Pedro L Baldoni, et al. Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. *Nature Methods*, 20(11):1810–1821, 2023.
32. Delaney K Sullivan and Lior Pachter. Flexible parsing, interpretation, and editing of technical sequences with splitcode. *Bioinformatics*, 40(6), 2024.
33. Ying Chen, Nadia M Davidson, Yuk Kei Wan, Harshil Patel, Fei Yao, Hwee Meng Low, Christopher Hendra, et al. A systematic benchmark of nanopore long read RNA sequencing for transcript level analysis in human cell lines. *bioRxiv*, 2021. doi: 10.1101/2021.04.21.440736.
34. Szi Kay Leung, Aaron R Jeffries, Isabel Castanho, Ben T Jordan, Karen Moore, Jonathan P Davies, Emma L Dempster, et al. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports*, 37(7):110022, 2021.
35. Isabel Castanho, Tracey K Murray, Eilis Hannon, Aaron Jeffries, Emma Walker, Emma Laing, Hedley Baulf, et al. Transcriptional signatures of tau and amyloid neuropathology. *Cell Reports*, 30(6):2040–2054.e5, 2020.
36. Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756, 2015.
37. Max Bloomfield, Samantha Hutton, Charles Velasco, Megan Burton, Miles Benton, Matt Storey, and ESR Genomics Consortium. Oxford nanopore next generation sequencing in a front-line clinical microbiology laboratory without on-site bioinformaticians. *Pathology*, 56 (3):444–447, 2024.
38. Oliver Cheng, Min Hao Ling, Changqing Wang, Shuyi Wu, Matthew E Ritchie, Jonathan Göke, Noorul Amin, and Nadia M Davidson. Flexiplex: a versatile demultiplexer and search tool for omics data. *Bioinformatics*, 40(3), 2024. doi: 10.1093/bioinformatics/btae102.
39. Laura Luebbert, Delaney K Sullivan, Maria Carilli, Kristján Eldjárn Hjörleifsson, Alexander Viloria Winnett, Tara Chari, and Lior Pachter. Efficient and accurate detection of viral sequences at single-cell resolution reveals novel viruses perturbing host gene expression. *bioRxiv*, 2023.
40. Kristoffer Sahlin and Veli Mäkinen. Accurate spliced alignment of long rna sequencing reads. *Bioinformatics*, 37(24):4643–4651, 2021.
41. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
42. Kristoffer Sahlin. Effective sequence similarity detection with strobemers. *Genome Research*, 31(11):2080–2094, 2021.
43. Rebekah K Loving, Delaney K Sullivan, Fairlie Reese, Elisabeth Rebboah, Jasmine Sakr, Narges Rezaie, Heidi Y Liang, Ghassan Filimban, Shimako Kawauchi, Conrad Oakes, et al. Long-read sequencing transcriptome quantification with lr-kallisto. *bioRxiv v1*, pages 2024–07, 2024.
44. Hyun Joo Ji and Mihaela Pertea. Enhancing transcriptome expression quantification through accurate assignment of long rna sequencing reads with transigner. *bioRxiv v2*, pages 2024–08, 2024.

Methods

lr-kallisto. Many approaches have been applied to RNA-seq quantification from classical alignment approaches to pseudoalignment paired with likelihoods and expectation-maximization (EM). Due to its speed, efficiency, and accuracy, pseudoalignment with likelihoods and EM has been widely adopted for the mapping of short read RNA-seq. However, for long-read RNA-seq, minimap2 has become the standard for aligning long-reads. Minimap2 follows the standard genome alignment methodology of seed-chain-align (20). It creates a reference index in the form of hashing minimizers into keys for a reference hash table storing the list of genomic/transcriptomic locations of the minimizer. For each read, minimap2 uses read minimizers as seeds matching these to the reference hash table and identifies the optimal collinear chain(s) of matches. While this method is accurate and has been developed to be highly efficient for the alignment strategy used, it is still time and resource expensive with high memory storage demands.

lr-kallisto, building on the existing framework of kallisto and adapting the pseudoalignment and expectation-maximization algorithm for long-reads, gives an accurate, fast, and low resource solution for mapping long-reads. The main technical challenge of long-reads lies in the higher sequencing error rates, though others include the differing rates of substitutions, deletions, and insertions between long-read sequencing technologies, sequencing length, repetitive regions, and concatemers. To address the challenge of higher sequencing error, different methods, including minimap2 (20), uLTRA (40), and STAR (41) have utilized various approaches to long-read alignment. Minimap2 uses a small k-mer size of 14 and 15 for long-reads, while uLTRA employs a two-pass chaining algorithm to improve alignment accuracy. Strobemers have been suggested using fuzzy k-mers that allow error tolerance (42). In lr-kallisto, we, instead, propose a long k-mer length and "chaining" pseudoalignment for addressing the challenges of long-read alignment.

We must address two points: first, that sequencing length and long-read sequencing error rates require a different algorithmic approach to pseudoalignment and, second, the length bias in sampling longer transcripts less times. To address the first, we propose the following algorithm for pseudoalignment and the change of index k-mer length to 63, which we discuss after describing the algorithm. Both of these changes take into consideration the sequencing error rate and repetitive regions across genes. While this idea is not a direct implementation of the chaining described in (20), it can be understood in a similar way. Within kallisto's pseudoalignment, a read's transcript compatibility class is determined. For short reads, this is accomplished with a strategy that increases efficiency by checking the transcript compatibility class for the first, middle, and end of k-mers in the read if the distance to the end of the contig is longer than the read or the first, middle, and end k-mers of the read within the region that is consistent with the contig in the transcriptome de Bruijn graph (T-DBG) (to ensure that the read is consistent with the T-DBG junctions) and then proceeds to the next contig in the read. If they are all the same, these are the only k-mers checked, while if they differ a more iterative approach is taken. We then take the intersection of these transcript compatibility classes. Whereas, in lr-kallisto, if the intersection of transcript compatibility classes (TCCs) a read maps to is empty, we instead take the most often occurring TCC. Moreover, if at least one k-mer maps uniquely to a transcript, then we take the most often occurring TCC among mapping k-mers that are uniquely mapping to a single transcript. In the case of the intersection, the intersection can directly be interpreted as the transcript or set of transcripts that the read has the longest combined stretches of compatibility with, since the intersection takes the subset of transcripts that coexist between all k-mers with compatible transcripts. However, the intersection may be empty in the case of a variant or error creating an isolated stretch of compatibility with a disjoint transcript compatibility class. Furthermore, in the case that the intersection is empty and the algorithm switches to using the mode of transcript compatibility classes with threshold one, the calculated mode across all transcript compatibility classes that k-mers in the read mapped to is the transcript or set of transcripts that again is the "longest chain" of compatibility.

The change of k-mer length to 63 was based on empirical evidence showing improved performance over the standard k-mer length of 31 for short reads. We found that across long-read technologies and simulations there was an improvement in metrics of Normalized Root Mean Squared Error and Spearman's correlation between lr-kallisto quantifications and the ground truth. In real data (both PacBio and ONT), we observed an increased rate of alignment of reads with a longer k-mer length for PacBio sequencing error rate less than 2% and for ONT sequencing error rate less than 10%. Moreover, the longer k-mer length improves the quality of mapping k-mers making it more probable that the read originates from the transcript compatibility class it maps to. As k increases, the number of distinct k-mers also increases, but the number of contigs decreases. This implies that the number of transcripts in a transcript compatibility class decreases on average with increasing length of k. Overall, the complexity of the T-DBG decreases (Supplementary Fig. 5), increasing the probability of the read originating from the transcript compatibility class it is mapping to. Furthermore, this also increases the probability of the intersection of equivalence classes being nonempty, which increases the overall mapping rate.

To address the second point, we adapted the effective length within kallisto to be transcript specific, i.e., defining the effective transcript length for a transcript t to be:

$$l_{\text{eff},t} = \frac{\sum l_{\text{reads aligning to } t}}{\#\text{ reads aligning to } t} - k.$$

We use the first 1 million aligning reads to compute these effective, transcript-specific lengths. We found that length normalization was effective at low sequencing error rates ($< 2\%$), providing a slight improvement in results, and was detrimental

to performance at high sequencing error rates.

Finally, we implemented a change to the expectation-maximization (EM) algorithm for long-reads. In the default option, we initialize transcript abundances to a uniform distribution on the multi mapping counts with the unique counts for each transcript added to the initialization of a transcript abundance. In the long-read option, we first apportion multi-mapping reads using the EM algorithm starting with a uniform distribution of multi-mapping reads among those mapped to transcripts, and then post EM we add the uniquely mapping counts to each transcript. We found that the latter option works better for the PacBio InDel profile with uniform error in reads and has reduced performance in the wild with real PacBio and ONT reads and simulations based on real data.

Mice and Tissue Collection. Mice were housed at the UC Irvine Transgenic Mouse Facility (TMF) in a temperature-controlled pathogen-free room under 12-hour light/dark cycles (lights on at 07:00 hr, off at 19:00 hr). The animal experiments were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC), protocol AUP-21-106, "Mouse genomic variation at single cell resolution". Left cerebral cortex tissues of 10-week-old mice were harvested from 4 C57BL/6J and 4 CAST/EiJ (2 males and 2 females per genotype) between the hours of 09:00 to 13:00. Tissues were stored in 1 mL Bambanker media in cryotubes kept at -80°C until nuclei isolation.

Purification of Nuclei from Mouse Tissues. Tissues were thawed in Bambanker media on ice until the tissue could be extracted and lysed using Nuclei Extraction Buffer (Miltenyi Biotec cat. #130-128-024). Using forceps, tissues were transferred to a chilled gentle MACS C Tube (Miltenyi Biotec cat. #130-093-237) with 2 mL Nuclei Extraction Buffer supplemented with 0.2 U/µL RNase Inhibitor (New England Biolabs cat. M0314L). Nuclei were dissociated from whole tissue using a gentleMACS Octo Dissociator (Miltenyi Biotec cat. #130-095-937). The resulting suspension was filtered through a 70 µm MACS SmartStrainer then a 30 µm strainer (Miltenyi Biotec cat. #130-110-916 and #130-098-458, respectively). Nuclei were resuspended in 3 mL PBS + 7.5% BSA (Life Technologies cat. #15260037) and 0.2 U/µL RNase inhibitor for manual counting using a hemocytometer and DAPI stain (Thermo Fisher cat. #R37606).

Nuclei Fixation. After counting, 4 million nuclei per sample were fixed using Parse Biosciences' Nuclei Fixation Kit v2 (cat. #ECF2003), following the manufacturer's protocol. Briefly, nuclei were incubated in fixation solution for 10 minutes on ice, followed by permeabilization for 3 minutes on ice. The reaction was quenched, then nuclei were centrifuged and resuspended in 300 µL Nuclei Buffer (Parse Biosciences cat. #ECF2003) for a final count. DMSO (Parse Biosciences cat. #ECF2003) was added before freezing fixed nuclei at -80°C in a Mr. Frosty (Sigma-Aldrich cat. #635639).

Split-Seq Experimental Protocol. Nuclei were barcoded using Parse Biosciences' WT Kit v2 (cat. #ECW02030), following the manufacturer's protocol. Fixed, frozen nuclei were thawed in a 37°C water bath and added to the Round 1 reverse transcription barcoding plate at 19,500 nuclei per well, with alternating columns in rows A and C containing C57BL/6J males and females and rows B and D containing CAST/EiJ males and females. In situ reverse transcription (RT) and annealing of barcode 1 + linker was performed using a thermocycler (Bio-Rad T100, cat. #1861096). After RT, nuclei were pooled and distributed in 96 wells of the Round 2 ligation barcoding plate for the in situ barcode 2 + linker ligation. After Round 2 ligation, nuclei were pooled and redistributed into 96 wells of the Round 3 ligation barcoding plate for the in situ barcode 3 + UMI + Illumina adapter ligation. Finally, nuclei were counted using a hemocytometer and distributed into 8 subpools of 13,000 nuclei. The nuclei in each subpool were lysed and cDNA was purified using AMPure XP beads (Beckman Coulter cat. #A63881), then the barcoded cDNA underwent template switching and amplification. Importantly, for two subpools ("13G" and "13H") we increased the number of PCR cycles to 13 cycles from 12, and increased the extension time from 3 minutes to 13 minutes in order to increase the yield of full-length barcoded cDNA. cDNA from one of the subpools ("13G") also received exome capture treatment using Parse Biosciences' Custom Gene Capture Kit (cat. #GCE1001) and a Mouse Exome Panel (Twist Bioscience, cat. #102036). 1 µg of cDNA was hybridized with a blocker solution to block repetitive sequences, then hybridized with the exome panel overnight. Captured molecules were purified using Streptavidin beads, then amplified again using the cDNA amplification reagents from the WT Kit v2 (Parse Biosciences cat. #ECW02030). The cDNA for all 8 subpools were cleaned using AMPure XP beads and quality checked using an Agilent Bioanalyzer before proceeding to Illumina and Nanopore library preparation. All 8 subpools were fragmented, size-selected using AMPure XP beads, and Illumina adapters were ligated. The cDNA fragments were cleaned again using beads and amplified, adding the fourth barcode and P5/P7 adapters, followed by size selection and quality checking with a Bioanalyzer. Libraries were sequenced with two runs of the Illumina NextSeq 2000 sequencer with P3 200 cycles kits (1.1 billion reads) and paired-end run configuration 140/86/6/0. Libraries with 5% PhiX spike-in were loaded at 1000 pM for one run and 1100 pM for the second run and sequenced to an average depth of # million reads per library.

Long-Read-Split-Seq Experimental Protocol and Base Calling. Nuclei were barcoded and cDNA was purified as specified in the previous section. LR-Split-seq libraries were generated using an input of 200 fmol from the amplified, barcoded

Split-seq cDNA before fragmentation (section 2 of the Split-seq protocol). Libraries were built using Oxford Nanopore Technologies Ligation Sequencing Kit (SQK-LSK114) and NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing (E7180L). The Short Fragment Buffer (SFB) from the Ligation Sequencing Kit (SQK-LSK114) during the second wash step. Libraries were loaded on R10.4.1 flowcells (FLO-PRO114M, FLO-MIN114) with an input of 20 fmol and 12 fmol, respectively. Sequencing was performed on the GridION and PromethION 2 Solo instruments using the MinKNOW software.

Bases were called from reads with Oxford Nanopore base-calling software Dorado v0.5.0 (https://github.com/nanoporetech/dorado) in super-accurate mode using config file dna_r10.4.1_e8.2_400bps_sup@v4.1.0 for both the exome capture and non-exome capture data, as well as the MinION and PromethION data.

Long-Read-Split-Seq Preprocessing and Quantification with splitcode and lr-kallisto. We first used splitcode to find barcodes and umis using linkers and reverse complements of linkers, allowing a total of 3 errors in linkers. We then used a custom python script to reverse the order of barcodes extracted from reverse strand to be in the same order as forward strand barcodes. Subsequently, we apply splitcode to combine and split randO and polyT barcodes from round 1 of Split-Seq barcoding, allowing 1 substitution or indel per barcode, 39,027,314 out of 105,591,654 raw reads passed this pipeline. We then use lr-kallisto to pseudoalign and quantify the resulting reads; 22,197,716 of the reads pseudoalign. We performed QC with a 500 UMI threshold per nuclei and filtered to genes present in at least 100 cells.

Error rate estimation. Error rates for the PacBio dataset (34) were calculated by analyzing a subsample of 1/8th of the reads using the NanoSim read characterization module with the command 'read_analysis.py transcriptome -i *fastq* -rg references/genome.fa -rt references/transcriptome.fa -annot references/annotations.gtf -t 8 -o output_folder'. Error rates for the LRGASP datasets were also calculated this way, without need for subsampling.

Benchmarking and comparisons. In benchmarks and comparisons of programs, we used Bambu v3.4.1, IsoQuant v3.3.0, and Oarfish v0.5.1. For the HCT116 data we also ran Oarfish 0.3.1 so as to be able to make a direct comparison with the results of (16). We ran Oarfish according to the scripts at https://github.com/COMBINE-lab/lr_quant_benchmarks/blob/0b89465420250d3511044fdc3d988a320aba73c6/snakemake_rules/isoquant_sim_data/alignment/alignment_transcriptome/align.snk and https://github.com/COMBINE-lab/lr_quant_benchmarks/blob/0b89465420250d3511044fdc3d988a320aba73c6/snakemake_rules/isoquant_sim_data/quantification/oarfish_quant/quant.snk. In a previous version of this preprint (43), Oarfish v0.3.1 and v0.4.0 were used and the simulation data was run with SAMtools sort as in (44). This appears to have resulted in overcounting that degraded Oarfish's performance.